# Outcome assessment in Ankylosing Spondylitis in focus

# Outcome assessment in Ankylosing Spondylitis in focus

ISBN: 90 5278 453 1

Cover design: Annett Gericke
Layout: Tiny Wouters
Production: Datawyse | Universitaire Pers Maastricht

The printing of this thesis was supported by Abbott BV, AstraZeneca BV, Bio-imaging Technologies BV, MSD, NV Organon, Schering-Plough and Wyeth Pharmaceuticals.

# Outcome assessment in Ankylosing Spondylitis in focus

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus, Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op vrijdag 22 april 2005 om 14.00 uur

door

Astrid Jacqueline Bernadette Wanders

UM
UNIVERSITAIRE PERS MAASTRICHT

Promotores

Prof. dr. D.M.F.M. van der Heijde
Prof. dr. Sj. van der Linden

Co-promotor

Dr. R.B.M. Landewé

Beoordelingscommissie

Prof. dr. C.P. van Schayck (voorzitter)
Prof. dr. R.A. de Bie
Prof. dr. B.A.C. Dijkmans (Vrije Universiteit Amsterdam)
Prof. dr. H. Mielants (Universiteit Gent, België)
Prof. dr. C.D.A. Stehouwer

Er bestaan geen feiten, alleen interpretaties

*Friedrich Nietzsche*

## Contents

# List of abbreviations

| | |
|---|---|
| ALAT | alanine aminotransferase |
| ANOVA | analysis of variance |
| AP | anteroposterior |
| AS | ankylosing spondylitis |
| ARR | absolute risk reduction |
| ASAS | assessment in ankylosing spondylitis |
| ASAT | aspartate aminotransferase |
| AUC | area under the curve |
| BASDAI | bath ankylosing spondylitis disease activity index |
| BASFI | bath ankylosing spondylitis functional index |
| BASMI | bath ankyloding spondylitis metrology index |
| BASRI | bath ankylosing spondylitis radiology index |
| COX | cyclo-oxygenase |
| DC-ART | disease controlling antirheumatic therapy |
| DMARD | disease modifying anti-rheumatic drug |
| DFI | dougados functional index |
| ES | effect size |
| ESR | erythrocyte sedimentation rate |
| HAQ | health appraisal questionnaire |
| ICC | interclass correlation coefficient |
| LR | likelihood ratio |
| MRI | magnetic resonance imaging |
| mSASSS | modified stoke ankylosing spondylitis spine score |
| mSv | milisievert |
| NPV | negative predictive value |
| NSAID | non-steroidal anti-inflammatory drug |
| NY | new york |
| OASIS | outcome ankylosing spondylitis international study |
| OMERACT | outcome measures in rheumatology |
| PPV | positive predictive value |
| RA | rheumatoid arthritis |
| RCT | randomized clinical trial |
| ROC | receiver operating characteristic |
| RRR | relative risk reduction |
| SASSS | stoke ankylosing spondylitis spine score |
| SAE | serious adverse event |
| SD | standard deviation |
| SF-36 | medical outcomes study short-form health survey |
| SI | sacroiliac |
| SMARD | symptom modifying antirheumatic drugs |
| SRM | standardized response mean |
| TNF | tumor-necrosis-factor |
| VAS | visual analogue scale |

# Chapter 1

# Introduction

# Ankylosing Spondylitis

On the cover of this thesis three radiographs are depicted, showing at two-year intervals the progressive abnormalities of the cervical spine of a young man who suffers from ankylosing spondylitis (AS). This condition is a chronic inflammatory rheumatic disease of uncertain etiology that primarily affects the axial skeleton (sacroiliac joints and spine). The term "ankylosing spondylitis" is derived from the Greek roots *ankylos* (bent), although now it usually implies fusion or adhesions, and *spondylos* (vertebral disc). Both inflammation and a combination of destructive and proliferative structural changes are features of the pathophysiological process. As a consequence of new bone formation vertebrae may become fused, which in turn may lead to impaired mobility of the spine. The radiographs on the cover show such new bone formation. At first glance, the first film presents a normal cervical spine, but a closer look reveals a fusion of the lower cervical vertebrae. The second film was taken two years later, and shows that the ankylosing process has progressed. The last film was taken 4 years after the first film, and shows almost complete ankylosis of the cervical spine.

Fortunately, not all patients with AS, whose total number is substantial with a worldwide prevalence ranging up to 0.9 % in Northern European Caucasian populations[1], experience such a rapid deterioration. The course of AS is highly variable and can be characterized by spontaneous remissions and exacerbations. The disease may be relatively mild, and limited to the sacroiliac joints in some patients, whereas it may be rapidly progressive, and associated with extra spinal manifestations such as arthritis, tendinitis, and acute uveitis anterior in other patients.

The basic treatment of AS consists of education, physiotherapy and exercises, and medication. Until recently drug therapy for AS was limited to non-steroidal anti-inflammatory drugs (NSAIDs). NSAIDs may rapidly relieve inflammatory back pain and reduce morning stiffness[2]. However, in order to experience sufficient clinical benefit, patients must take NSAIDs regularly, in full anti-inflammatory dose. The clinical benefit of these drugs does not persist after discontinuation. Furthermore, NSAID use is limited by the increased risk of serious gastro-intestinal side effects such as ulcers, perforations and bleedings. These side effects are due to the inhibition of cyclo-oxygenase -1 (COX-1). The new generation of NSAIDs selectively inhibits COX-2 that is up regulated under inflammatory conditions and responsible for the production of pro-inflammatory prostaglandins. Selective COX-2 inhibitors leave COX-1 relatively undisturbed. COX-2 selective NSAIDs are associated with a reduced risk of the above mentioned serious gastro-intestinal complications[3,4] and are at least as effective in ankylosing spondylitis as conventional NSAIDs[5,6].

A new era of drug treatment in AS has started with the introduction of tumor-necrosis-factor inhibiting (anti-TNF) therapy. After the impressive results in the treatment in rheumatoid arthritis (RA), the clinical efficacy of TNF-blocking drugs was also investigated in AS. Randomized double blind, placebo controlled clinical trials have

proven the efficacy of infliximab[7], and etanercept[8] in reducing disease activity, assessed by clinical measures and by acute phase reactants, as well as improving spinal mobility in patients with AS.

## Outcome assessment

These new therapeutic opportunities have further stimulated research in AS. It is pivotal that all studies use the same outcome parameters, which will allow an appropriate comparison of results across studies, which in turn will eventually lead to better informed health care providers and evidence-based treatment options for patients. In 1995, the international ASsessment in Ankylosing Spondylitis (ASAS) working group was formed. Its aim was to develop internationally standardized endpoints for use in clinical trials as well as in clinical practice. Core sets have been developed by ASAS for the following three settings: symptom modifying antirheumatic drugs (SMARD)/physical therapy, clinical record keeping and disease controlling antirheumatic therapy (DC-ART)[9]. The domains physical function, pain, spinal mobility, spinal stiffness, fatigue and patient's global assessment are included in all three settings. In addition, the domains peripheral joints and entheses, and acute phase reactants have been added for the settings DC-ART and clinical record keeping, and the domain radiographic assessment for the DC-ART setting (figure 1.1). Selection of the specific instruments to be included in each domain was determined by consensus among ASAS members[10].

## ASAS core sets

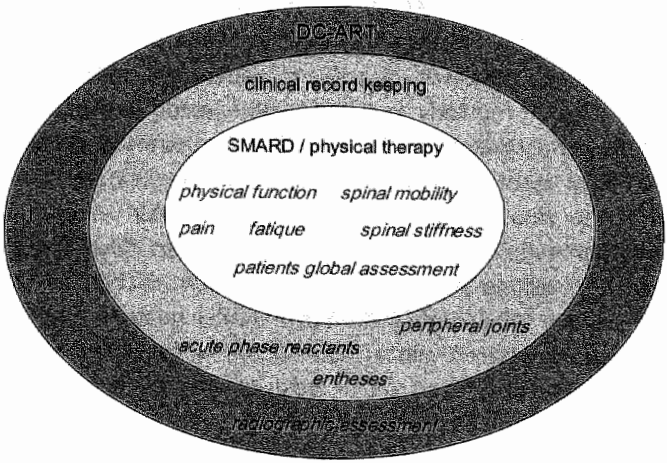



Figure 1.1 van der Heijde et al, J Rheumatol, 1997. 24(11): p. 2225-9; updated ASAS workshop Gent Oct 2002

# This Thesis

This thesis focuses on outcome assessment in AS. Since recent studies have demonstrated that anti –TNF therapy is highly efficacious in relieving signs and symptoms and improving physical function of patients with AS, anti-TNF therapy could be considered the first real disease controlling therapy if an effect on preventing structural damage could be proven. With these new developments research focusing on the DC-ART core set of the ASAS has become very relevant.

The data set of one of the clinical trials with anti-TNF therapy (etanercept) offered the opportunity to evaluate the responsiveness and discriminative capacity of the various measures of the ASAS DC-ART core set (chapter two). In order to detect changes after effective treatment, a measure should be responsive. Highly responsive measures are preferred because they facilitate the detection of improvement. However, responsiveness alone is not sufficient to assure the detection of small – but potentially relevant – differences between effective treatment and placebo or two effective treatments. Discriminative capacity – the ability of an assessment to distinguish between active treatment and placebo, or between two active treatments – should also be acceptable. The purpose of the study was twofold. First, to investigate the responsiveness and discriminative capacity of the ASAS DC-ART core set in a trial with etanercept in AS[8]. Second, to evaluate the relative importance of the different measures included in the ASAS DC-ART core set with respect to responsiveness and discriminatory capacity.

All domains of the ASAS-DCART core set were evaluated in the above mentioned trial, except radiographic assessment because of the short follow-up of the trial. However, a short follow-up is not the only limiting step in the assessment of radiographic progression in clinical trials with anti-TNF therapy in AS. The ASAS working group has selected radiographic assessment as a domain to be assessed in potential disease controlling drugs but did not select the instrument for this domain because of lack of appropriate data. The selection of a proper instrument to assess progression of structural damage, however, has become very relevant in light of the experience in RA with respect to the anti-TNF-induced retardation of progression of joint damage[11]. So, in the field of AS in the year 2005, one of the most pressing questions is whether or not anti-TNF therapy is associated with an arrest or at least retardation of structural damage measured on radiographs of the spine. Therefore the focus of this thesis is on radiographic assessment of AS.
In chapter 3 the available scoring methods for radiographic assessment in AS are described and different aspects of validity are compared. The investigated methods are the Bath Ankylosing Spondylitis Radiology Index (BASRI)[12], the Stoke Ankylosing Spondylitis Spine Score (SASSS)[13] and a modification of the SASSS, the so-called modified SASSS[14]. In earlier research, independent observers have compared reliability

and sensitivity to change of these methods[15,16]. However these comparisons have limitations. First, they were only partly based on the OMERACT filter[17], which is proposed in rheumatology research for a proper selection of a measure. Apart from reliability and sensitivity to change, captured as the aspect discrimination in the OMERACT filter, there are two other aspects; truth and feasibility, which have not yet been evaluated. Second, the earlier investigated comparisons of reliability concerned *status* scores. However, in a clinical trial the change between two moments: the *progression* score is of more importance than the score at one point in time: the *status* score. Reliability of status scores and progression scores may differ importantly. To abolish these limitations the study presented in this chapter was performed. The aim was to test the radiographic scoring methods for all three aspects of the OMERACT filter including reliability of progression scores.

Another important aspect in the evaluation of progression of radiographic damage in clinical trials is the way in which films are presented to observers. It is known from studies concerning evaluation of radiographic damage in RA that the order in which films are presented to the observer may influence the results[18-21]. Films can be grouped per patient and presented to the reader without awareness of the chronological order of the films: paired scoring. Films can also be grouped per patient and presented in chronological order. The issue which of both reading orders should be used has not yet been investigated in radiographic assessment in AS. The reading of structural damage in RA clinical trials is predominantly performed by readers blinded for the sequence. Therefore it seems obvious that radiographic progression in AS clinical trials should also be assessed by paired reading. However the disease course concerning radiographic progression in AS differs from RA. Compared to RA progression in AS occurs slowly and, maybe more relevant, progression seems to occur only in a minority of patients. So, there is some concern that paired scoring in AS is not sensitive enough. In chapter four a study with a twofold aim is described: 1) to explore the differences with respect to sensitivity to change between paired and chronological scoring in AS, and 2) to investigate whether trials with radiographic progression as primary endpoint can be designed, that have sufficient statistical power with feasible patient numbers if films are read with paired order.

Issues concerning radiographic assessment were discussed in the above mentioned chapters, 2 to 4. It might be clear that radiographic evaluation is a complex and therefore time-consuming process. This is very useful for the evaluation of efficacy of drugs in clinical trials. However, in general practice it would be advantageous if the same amount of relevant information could be obtained in a more feasible way. A disadvantage of radiographs for patients is the fact that radiographs are associated with radiation exposure and for the health care system the costs related to radiographic assessment. So it might be useful to investigate whether there is another cheaper, safer and less time-consuming assessment that can serve as a proxy for radiographic

assessment in individual patients. Radiographic damage, especially if fusion of the vertebrae is involved, leads to impaired spinal mobility. It is also known that on a group level there is a clear association between radiographic damage and reduced spinal mobility. In chapter 5 we evaluated the relationship between various spinal mobility assessments and radiographic damage on an individual patient level and assessed the appropriateness of using these spinal mobility measurements as a proxy for structural damage assessed on radiographs.

In chapter 5 the focus is on the individual patient. In chapter 6 the focus is again on groups of patients, namely patients in clinical trials. It is likely that the risk that patients have on a particular radiographic outcome is co-determined by the radiographic status at baseline[22]. If so, the selection of patients may influence the radiographic outcome in clinical trials. This is described in chapter 6. The data described in that study are derived from clinical trials in RA because in contrast to clinical trials in AS, in clinical trials in RA radiographic outcome already has a prominent place. However, the general concept which forms the basis of the hypothesis in chapter 6 is also applicable to trial design in AS[22].

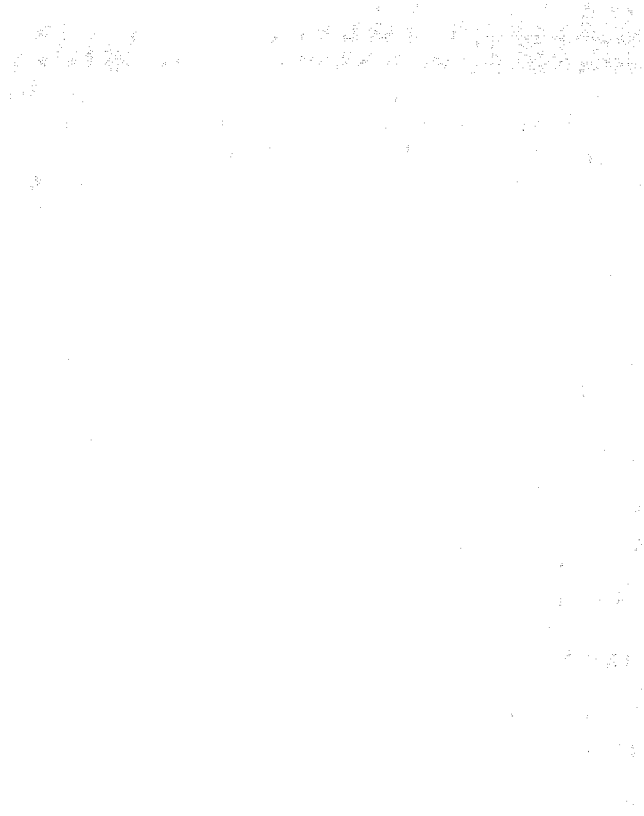
After studies that describe methodological issues and a brief excursion to RA, in chapter 7 the acquired knowledge is now applied: assessment of radiographic progression in a clinical trial. Because of the risk of chronic NSAID-use, many rheumatologists advise to use NSAIDs only when needed to actually reduce signs and symptoms. So the long-term effects of chronic NSAID-use are unknown. The improved safety profile of COX-2 selective- as compared to unselective - NSAIDs justifies a formal test of the hypothesis that NSAIDs may alter the course of AS. This hypothesis was tested in a unique trial design. Patients participating in a 6-week placebo controlled trial comparing two NSAIDs, were randomized into two groups at the final visit. Two treatment strategies were compared: long-term continuous NSAID-use versus NSAID-use on demand only. After a total follow-up of two years the efficacy of the two treatment strategies were compared with as primary outcome radiographic progression blindly assessed by the modified SASSS. This was the first randomized trial in AS of any drug with radiographic progression as the primary endpoint.

Chapter 8 consists of a summary and general discussion on the findings of this thesis. A summary of this thesis in Dutch is provided in Chapter 9.

# References


1. Braun J, Bollow M, Remlinger G, Eggens U, Rudwaleit M, Distler A, Sieper J. Prevalence of spondylarthropathies in HLA-B27 positive and negative blood donors. Arthritis Rheum 1998;41:58-67.
2. Dougados M, Revel M, Khan MA. Spondylarthropathy treatment: progress in medical treatment, physical therapy and rehabilitation. Baillieres Clin Rheumatol 1998;12:717-36.
3. Langman MJ, Jensen DM, Watson DJ, Harper SE, Zhao PL, Quan H, Bolognese JA, Simon TJ. Adverse upper gastro-intestinal effects of rofecoxib compared with NSAIDs. JAMA 1999;282:1929-33.
4. Simon LS, Weaver AL, Graham DY, Kivitz AJ, Lipsky PE, Hubbard RC, Isakson PC, Verburg KM, Yu SS, Zhao WW, Geis GS. Anti-inflammatory and upper gastro-intestinal effects of celecoxib in rheumatoid arthritis: a randomized controlled trial. JAMA 1999;282:1921-8.
5. Dougados M, Behier JM, Jolchine I, Calin A, van der Heijde DMFM, Olivieri I, Zeidler H, Herman H. Efficacy of celecoxib, a cyclooxygenase 2-specific inhibitor, in the treatment of ankylosing spondylitis: a six-week controlled study with comparison against placebo and against a conventional nonsteroidal antiinflammatory drug. Arthritis Rheum 2001;44:180-5.
6 Melian A, van der Heijde DM, James MK, Calin A, Giallella KM, Reicin AS, et al. Etoricoxib in the treatment of ankylosing spondylitis. ACR annual scientific meeting 2002 abstract 1131
7. Braun J, Brandt J, Listing J, Zink A, Alten R, Golder W, Gromnica-Ihle E, Kellner H, Krause A, Schneider M, Sorensen H, Zeidler H, Thriene W, Sieper J. Treatment of active ankylosing spondylitis with infliximab: a randomised controlled multicentre trial. Lancet 2002;359: 1187-93.
8. Gorman JD, Sack KE, Davis JC. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor alpha. N Engl J Med 2002;346:1349-56.
9. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. Assessments in Ankylosing Spondylitis Working Group. J Rheumatol 1997;24:2225-9.
10. van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. Assessments in Ankylosing Spondylitis. J Rheumatol 1999;26: 951-4.
11. Lipsky PE, van der Heijde DM, St Clair EW, Furst DE, Breedveld, FC, Kalden JR, Smolen JS, Weisman M, Emery P, Feldmann M, Harriman GR, Maini RN; Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. N Engl J Med 2000;343: 1594-602.
12. MacKay K, Mack CS, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI): a new, validated approach to disease assessment. Arthritis Rheum 1998;41: 2263-70.
13. Averns HL, Oxtoby HJ, Taylor HG, Jones PW, Dziedzic K, Dawes PT. Radiological outcome in ankylosing spondylitis: use of the Stoke Ankylosing Spondylitis Spine Score (SASSS). Br J Rheumatol 1996;35:373-6.

14. Creemers MCW, Franssen M, van 't Hof MA, Gribnau FWJ, van de Putte LBA, van Riel PLCM. A radiographic scoring system and identification of variables measuring structural damage in ankylosing spondylitis. [thesis], 1993. University of Nijmegen (The Netherlands). Published later as:
Creemers M, Franssen M, van 't Hof M, Gribnau F, van de Putte L, van Riel P. Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. Ann Rheum Dis, 2004. Published Online First [March 29,2004] doi.10.1136/ard.2004.020503.
15. Spoorenberg A, de Vlam K, van der Linden S, Dougados M, Mielants H, van der Tempel H, van der Heijde D. Radiological scoring methods in ankylosing spondylitis: reliability and sensitivity to change over one year. J Rheumatol 1999;26:997-1002.
16. Spoorenberg A, DeVlam F, van der Linden S, Dougados M , Mielants H, van der Tempel H, van der Heijde D. Radiological scoring methods in Ankylosing Spondylitis. Reliability and sensitivity to change over one and two years. J Rheumatol 2004;31:125-32.
17. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol 1998;25:198-9.
18. Ferrara R. Priolo F, Cammisssa M, Baccarini L, Cerase A, Pasero G et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the GRISAR Study. Gruppo Reumatologi Italiani Studio Artrite Reumatoide. Ann Rheum Dis 1997;56:608-12.
19. Salaffi F,Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. J Rheumatol 1997;24:2055-6.
20. van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology 1999;38:1213-20.
21. Bruynesteyn K, van der Heijde D, Boers M, Saudan A, Peloso P, Paulus H, Houben H, Griffiths B, Edmonds J, Bresnihan B, Boonen A, Van Der Linden S. Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. J Rheumatol 2002;29:2306-12.
22. van der Heijde D, Wanders A, Mielants H, Dougados M, Landewé R. Prediction of progression of radiographic damage over 4 years in patients with ankylosing spondylitis. Ann Rheum Dis 2004;63(S1): OP0132



# Chapter 2

# Responsiveness and discriminatory capacity of the ASAS DC-ART core set and other outcome measures in a trial of etanercept in ankylosing spondylitis

A Wanders, J Gorman, J Davis, R Landewé, D van der Heijde

*Arthritis Care & Research 2004, 51(1):1-8*

# Abstract

**Objective**
To investigate the responsiveness and discriminative capacity, and the relationship between both, of instruments selected for the disease controlling antirheumatic therapy (DC-ART) core set by the Assessments in Ankylosing Spondylitis (ASAS) Working Group.

**Methods**
Responsiveness and discriminative capacity of different measures reflecting disease activity and function, either included in the ASAS DC-ART core set or not, were evaluated in a randomized placebo-controlled clinical trial comparing etanercept with placebo in patients with AS, using Guyatt's method as the primary analysis for responsiveness, and Student's t-test for discriminative capacity.

**Results**
At day 28 of therapy, almost all measures indicated moderate to large responsiveness in the etanercept group (Guyatt 0.60-3.11). Some scales of the Short Form-36 (general health, mental component summary and role-emotional), the modified Schober's test, and the Fatigue Severity Scale were not responsive. The results were similar if analyzed at day 112 of therapy. Peripheral joint counts and –scores and occiput to wall distance could not be evaluated due to a floor-effect. In general, the relation between responsiveness and discriminative capacity was strong: measures that demonstrated high responsiveness also showed high between-group t-values.

**Conclusion**
Measures included in the ASAS DC-ART core set, except modified Schober's test, have good responsiveness and good discriminatory capacity. Some measures could not be evaluated due to a floor effect.

# Introduction

There are many outcome measures available to evaluate patients with ankylosing spondylitis (AS). In 1995, in order to develop internationally standardized endpoints for use in clinical trials as well as practice, the ASsessments in Ankylosing Spondylitis (ASAS) working group was formed. Core sets were developed by the ASAS for the following three settings: disease controlling antirheumatic therapy (DC-ART), symptom modifying antirheumatic drugs (SMARD)/physical therapy and clinical record keeping[1]. The domains physical function, pain, spinal mobility, stiffness, and the patient's global assessment are included in all three settings. In addition, the domains peripheral joints and entheses, and acute phase reactants are added for the settings DC-ART and clinical record keeping, and the domains X-ray (of spine and hips) and fatigue for the DC-ART setting. Selection of the specific instruments to be included in each domain was determined by consensus among ASAS members[2]. The instruments selected for the DC-ART core set have not yet been validated with respect to the OMERACT filter: truth, discrimination and feasibility[3]. Recent studies have suggested that tumor-necrosis-factor-alpha inhibiting (anti–TNF-alpha) therapy may be promising for ankylosing spondylitis. Anti-TNF alpha therapy has demonstrated to be efficacious in spondylitis[4,5], and therefore the data set of a clinical trial with anti-TNF alpha therapy offers an opportunity to evaluate the ASAS DC-ART core set. In order to discriminate in trials between effective treatment and placebo, a measure should be responsive. Highly responsive measures are preferred because they facilitate the detection of improvement. However, responsiveness alone is not sufficient to assure the detection of small - but potentially important – differences between effective treatment and placebo (discriminative capacity). The purpose of this study was twofold. First, to investigate the responsiveness and discriminatory capacity of the ASAS DC-ART core set in a trial with etanercept in AS[6]. Second, to evaluate the relationship between responsiveness and discriminatory capacity of the measures included in the ASAS DC-ART core set.

# Patients and methods

### The clinical trial with anti-TNF aplha Therapy

The study was a four-month double-blind clinical trial that randomly assigned 40 patients with AS (defined by the modified New York criteria[7] to the treatment or placebo group. All patients had an active spondylitis, defined by morning stiffness of more than 45 minutes, inflammatory back pain, patient and physician global assessment of disease activity of moderate or higher. Patients were allowed to continue standard therapies for AS as long as they were on a stable dose. Patients received 25 mg of etanercept or placebo subcutaneously twice weekly for four months.

DC-ART Core Set measures

For the seven DC-ART domains the following instruments were assessed:

| | |
|---|---|
| *Function:* | The BASFI[8] and the Dougados Functional Index (DFI)[9]. |
| *Pain:* | Visual analogue scale (VAS) for spinal pain at night over the past week and an overall VAS due to spinal pain over the past week. |
| *Spinal mobility:* | Chest expansion[10], modified Schober's test[11] and occiput-to-wall distance. |
| *Patient global:* | Assessment of patient's global well-being rated on a 5- point Likert scale over the past week. |
| *Stiffness:* | Duration of morning stiffness expressed in minutes experienced on the day preceding the visit. |
| *Peripheral joints and entheses:* | Number of swollen joints, counted in 42 diarthrodial joints. |
| *Acute phase reactants:* | The erythrocyte sedimentation rate (ESR) was measured as an acute phase reactant. |

Other single measures and indices

For additional assessments of the peripheral joints and entheses the following measures were added:
Joint tenderness was counted (tender joint count) and tenderness and swelling scored (tender joint score and swollen joint score) in 44 diarthrodial joints (44 joints for tenderness evaluation and 42 joints for swelling evaluation (no hips in swollen joint score/count): rated on a 4-point scale (0 = no swelling, 1 = mild (detectable synovial thickening without loss of bony contours), 2 = moderate (loss of distinctness of bony contours), 3 = severe (bulging synovial proliferation with cystic characteristics)). Enthesopathy was scored by means of the modified enthesopathy index: uniform manual pressure is applied to the vertebral processes of C1-C2, C7-Th1, Th12-L1, L5-S1, the symphysis pubis, both greater trochanters, pelvic abductor origin, anterior superior border of the iliac crests, ischial tuberosities, insertions of Achilles tendons, and plantar fascia and tenderness was scored on a 4 point scale (0 = no pain, 1 = mild tenderness, 2 = moderate tenderness, 3 = wince or withdrawal).
Physician global assessment was measured by means of a VAS of overall disease activity. Fatigue was measured with the Fatigue Severity Scale[12] and quality of life was assessed with the Medical Outcomes Study Short-Form Health Survey (SF-36)[13]. The SF-36 has 8 scales: physical functioning, social functioning, role limitations due to a physical problem, role limitations due to an emotional problem, mental health, vitality, pain and general health. There are two summary measures (physical and mental component summary) calculated from scores of the individual scales.

### Statistical analysis

All analyses were based on intention to treat: only 3 patients dropped out of the trial. A last-value-carried-forward approach was done for the data obtained from these patients. The day 28 and day 112 data are reported to allow further exploration of trends in responsiveness and discriminatory power.

Three statistical methods were used to assess responsiveness: the standardized response mean (SRM), the effect size (ES) and the Guyatt method. The Guyatt method is viewed by some as the superior responsiveness statistic, since this statistic takes into account the variability of the placebo group[14]. Ranking the measures in order of the Guyatt method was used as the primary responsiveness statistic in this study.

*Guyatt method*: The formula for Guyatt's responsiveness index[15] is: $\Delta_x / \sqrt{2} * MSE_x$ where $\Delta_x$ = minimally clinically important change on the measure and $MSE_x$ is the mean squared error of $X$ obtained from an analysis of variance (ANOVA) model that examines repeated observations of the measure in clinically stable subjects. Alternatively, if there are only two observations of the measure $MSE_x$ is the standard deviation of the individual change scores in clinically stable patients (i.e., placebo-treated patients)[16]. Guyatt's index is calculated as the ratio of the mean change of patients in the etanercept group divided by the standard deviation (SD) of change of patients in the placebo-group[17].

*Standardized response mean* (SRM): The SRM is calculated as the mean change in scores divided by the SD of these changes.

*Effect size*: The effect size is the difference between the mean baseline and follow-up scores on the measure, divided by the standard deviation of the baseline scores.

For all responsiveness statistics, values of 0.20, 0.50 and 0.80 or greater have been advocated to represent small, moderate, and large responsiveness, respectively[18-20].

*Assessment of floor effects:* Floor effects may impair responsiveness because patients with very low baseline values cannot improve further. Histograms of the cross-sectional analysis were used to examine the presence of floor effects.

*Discrimination:* The independent unpaired Student's t-test values of each variable are reported for the discriminatory capacity. All t-tests were 2-sided with a significance level ($\alpha$) of 0.05 resulting in a critical t-value of 2,03.

*Relationship responsiveness and discriminatory capacity:* To evaluate the relationship between responsiveness and discriminatory capacity, the responsiveness statistics are plotted versus the t-values. For the correlation between the primary responsive statistic (the Guyatt method) and the discriminatory capacity, the Spearman correlation coefficient is also calculated.

# Results

Both the treatment and placebo group included 20 patients[6]. The groups were acceptably balanced with respect to important demographic and prognostic variables (table 2.1).

Table 2.1 Baseline characteristics of patients

| Characteristics | Etanercept n=20 | Placebo n=20 |
|---|---|---|
| Male (%) | 65 | 90 |
| White (%) | 75 | 70 |
| HLA-B27 positive (%) | 95 | 90 |
| Age, mean ± SD (years) | 38 ± 10 | 39 ± 10 |
| Disease duration, men ± SD (years) | 15 ± 10 | 12 ± 9 |

Some outcome measures demonstrated an extremely skewed distribution by histograms and analyses for responsiveness and discriminatory capacity were not performed for these variables. The following measures were excluded from further analysis: swollen joint score, tender joint score, swollen joint count, tender joint count, modified enthesopathy index, and occiput-to-wall distance. However, in those patients with the potential for improvement in the above measures, a positive treatment effect of etanercept was seen (tables 2.2 and 2.3).

At day 28, most remaining measures indicated moderate to large responsiveness (Guyatt 0.60-3.11), except several components of the SF-36 (general health, mental component summary and role-emotional), the modified Schober and the Fatigue Severity Scale (table 2.2, figure 2.1). The results for day 112 were generally similar, with somewhat larger responsiveness for most measures (Guyatt 0.51–3.77), except some components of the SF-36 (mental component summary and role-emotional), and the modified Schober (table 2.3, figure 2.2). More measures demonstrated low responsiveness when results from all 3 statistical methods were considered (<0.50); chest expansion, SF-36: mental health, for day 28 and chest expansion, Fatigue Severity Scale, and SF-36: mental health for day 112.

Concerning the ASAS core set (table 2.4) it can be seen that for the domain *function*, both the BASFI and DFI demonstrated a large degree of responsiveness at day 112. At day 28, BASFI demonstrated a greater responsiveness than the DFI (large vs. moderate respectively). The BASFI also appeared to have a higher discriminative power than the DFI.

Both instruments in the domain *pain* showed excellent responsiveness statistics and a high discriminative power. The responsiveness for the instruments in the domain *spinal mobility* was only good for the outcome measure chest expansion. The modified

Schober had a small responsiveness and occiput-to-wall was not analyzed due to a floor-effect. In addition, both measures could not significantly discriminate between placebo and etanercept-treated patients.
The domain of *patient global assessment* has excellent responsiveness statistics and a high discriminative power.
The domain *stiffness* was largely responsive and highly discriminative within the first month of treatment, although the discriminative capacity decreased by day 112 and was only barely significant.
The number of swollen joints, the instrument for the domain *peripheral joints and entheses,* was susceptible to a floor effect, and so was not able to be evaluated for responsiveness. The instrument ESR for the final domain *acute phase reactants* has an excellent responsiveness and discriminatory capacity.
The results are presented at two time points, day 28 and day 112. It is noticeable that by 28 days some variables show already a good responsiveness (Guyatt>0,80) and discriminatory capacity (t>2,03). By the end of the trial, these measures still have\ a good responsiveness and discriminatory capacity.
The following variables belong to this group: ESR, physician global assessment, BASFI, VAS nocturnal, patient global assessment, VAS overall and 5 of the SF-36 scales (physical component summary, physical functioning, bodily pain, role physical and vitality).
When attention is paid to the rank order of the measures at the two time points, it can be seen those measuring similar outcomes appear to aggregate by the end of the trial. For instance, while at day 28 the SF-36: vitality scale was superior to the Fatigue Severity Scale in responsiveness and discriminatory power, these differences were lessened by the end of the trial.
The relationship between responsiveness statistics and between-group discrimination is seen in figure 2.1 and 2.2. In general, there appeared to be a strong correlation between responsiveness and discriminatory capacity. The Spearman correlation coefficients for the Guyatt responsiveness statistics and t-values are 0,92 and 0,85, respectively at day 28 and 112.

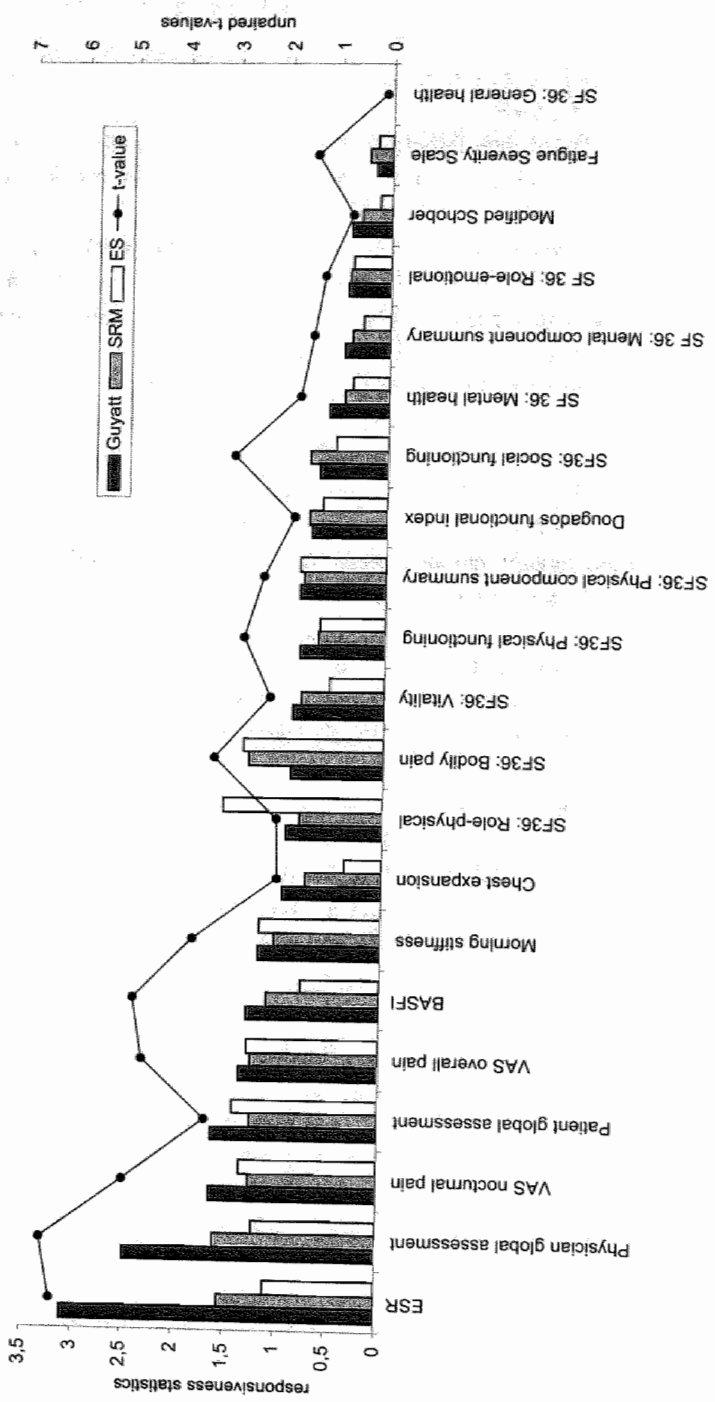



Figure 2.1 Comparison of responsiveness and discrimination performance of measures at day 28.

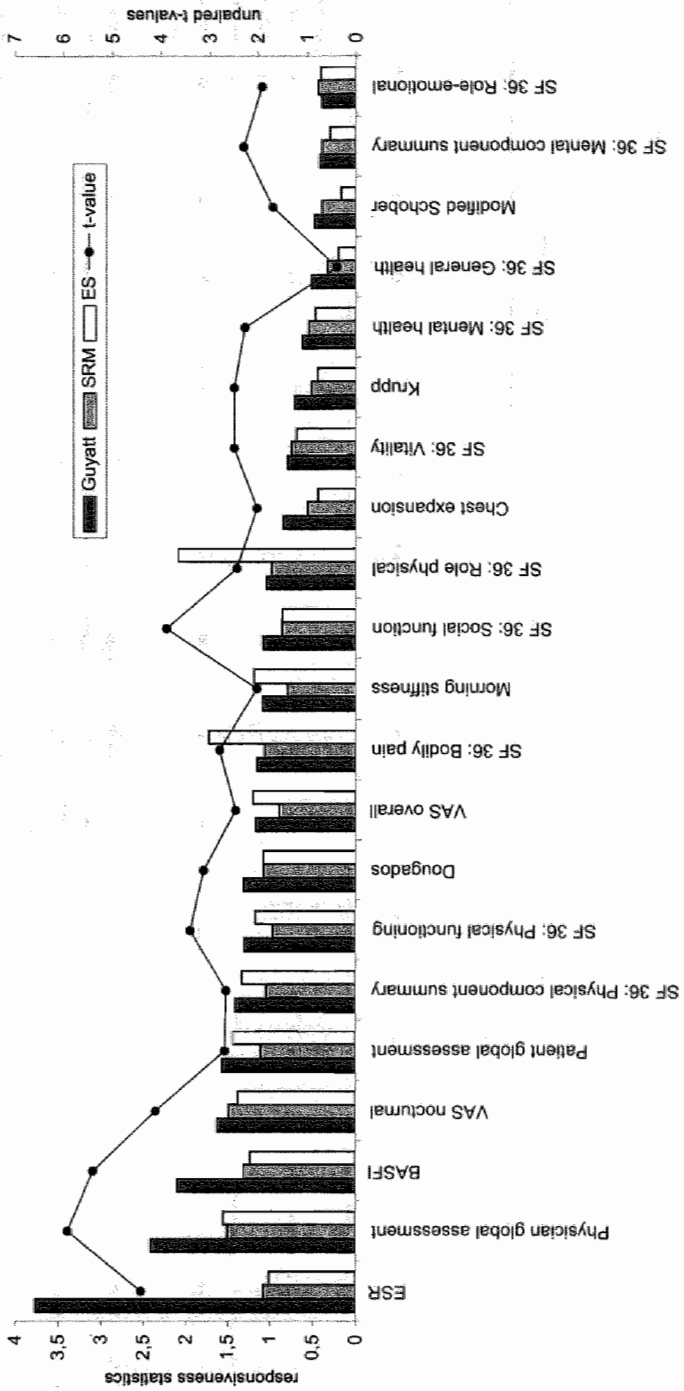



Figure 2.2 Comparison of responsiveness and discrimination performance of measures at day 112.

Table 2.2 Indices of responsiveness at day 28 of follow up, for the treatment and placebo group, ordered by the Guyatt method*

| | Treatment group (n=20) Mean change | SD | SD baseline | SRM | ES | Mean change | SD | SD baseline | SRM | ES | Guyatt | t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ESR | 25.35 | 16.31 | 23.09 | 1.55 | 1.10 | -0.85 | 8.16 | 16.33 | -0.10 | -0.05 | 3.11 | 6.42 (2) |
| Physician global assessment | 23.85 | 14.83 | 19.62 | 1.61 | 1.22 | -2.53 | 9.59 | 16.43 | -0.26 | -0.15 | 2.49 | 6.63 (1) |
| VAS nocturnal pain | 32.35 | 25.81 | 23.89 | 1.25 | 1.35 | -3.95 | 19.45 | 25.34 | -0.20 | -0.16 | 1.66 | 5.02 (3) |
| Patient global assessment | 1.05 | 0.83 | 0.73 | 1.27 | 1.44 | 0.25 | 0.64 | 0.69 | 0.39 | 0.36 | 1.64 | 3.43 (6) |
| VAS overall pain | 27.70 | 22.10 | 21.30 | 1.25 | 1.30 | -3.50 | 20.14 | 24.88 | -0.17 | -0.14 | 1.38 | 4.67 (5) |
| BASFI | 1.62 | 1.45 | 2.06 | 1.12 | 0.79 | -0.45 | 1.24 | 2.45 | -0.37 | -0.18 | 1.31 | 4.85 (7) |
| Morning stiffness | 60.00 | 56.99 | 50.56 | 1.05 | 1.19 | -3.00 | 49.83 | 70.69 | -0.06 | -0.04 | 1.20 | 3.72 (4) |
| Chest expansion | 0.60 | 0.79 | 1.63 | 0.76 | 0.37 | 0.14 | 0.60 | 1.65 | 0.22 | 0.08 | 0.99 | 2.08 (10) |
| SF-36: role physical | 34.98 | 42.42 | 22.16 | 0.82 | 1.58 | 8.75 | 36.52 | 43.13 | 0.24 | 0.20 | 0.96 | 2.10 (15) |
| SF-36: bodily pain | 20.95 | 15.81 | 15.20 | 1.33 | 1.38 | 0.00 | 23.04 | 21.84 | 0.00 | 0.00 | 0.91 | 3.35 (12) |
| SF-36: vitality | 10.23 | 12.35 | 18.83 | 0.83 | 0.54 | 1,75 | 11.39 | 16.75 | 0.15 | 0.10 | 0.90 | 2.26 (13) |
| SF-36: physical functioning | 12.00 | 18.02 | 18.52 | 0.67 | 0.65 | -2.25 | 14.09 | 23.37 | -0.16 | -0.10 | 0.85 | 2.79 (11) |
| SF-36: physical component summary | 6.91 | 8.50 | 8.20 | 0.81 | 0.84 | 0.53 | 8.14 | 10.09 | 0.06 | 0.05 | 0.85 | 2.42 (9) |

| | Treatment group (n=20) Mean change | SD | SD baseline | SRM | ES | Mean change | SD | SD baseline | SRM | ES | Guyatt | t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dougados functional index | 3.90 | 5.10 | 6.12 | 0.76 | 0.64 | 0.90 | 5.22 | 6.77 | 0.17 | 0.13 | 0.75 | 1.84 (8) |
| SF-36: social functioning | 13.10 | 16.92 | 25.22 | 0.77 | 0.52 | -4.38 | 19.57 | 25.74 | -0.22 | -0.17 | 0.67 | 3.02 (14) |
| SF-36: mental health | 7.20 | 16.50 | 19.88 | 0.44 | 0.36 | -0.80 | 11.94 | 15.71 | -0.07 | -0.05 | 0.60 | 1.76 (16) |
| SF-36: mental component summary | 3.44 | 9.16 | 12.97 | 0.38 | 0.27 | -0.62 | 7.63 | 9.24 | -0.08 | -0.07 | 0.45 | 1.52 (17) |
| SF-36: role emotional | 16.67 | 42.58 | 44.43 | 0.39 | 0.38 | 0.00 | 39.99 | 38.84 | 0.00 | 0.00 | 0.43 | 1.29 (20) |
| Modified Schober | 0.18 | 0.61 | 1.50 | 0.30 | 0.12 | 0.05 | 0.45 | 1.48 | 0.10 | 0.03 | 0.40 | 0.78 (18) |
| Fatigue Severity Scale | 0.23 | 1.03 | 1.55 | 0.22 | 0.15 | -0.35 | 1.39 | 1.52 | -0.25 | -0.23 | 0.16 | 1.49 (19) |
| SF-36: general health | -0.15 | 11.82 | 25.62 | -0.01 | -0.01 | -0.80 | 16.77 | 19.15 | -0.05 | -0.04 | -0.01 | 0.14 (21) |
| Modified enthesopathy index | 3.40 | 4.92 | 8.43 | † | † | 0.9 | 2.51 | 7.88 | † | † | † | † |
| Occiput-to-wall distance | 0.38 | 1.78 | 7.92 | † | † | -0.20 | 0.68 | 3.54 | † | † | † | † |
| Swollen joint count | 0.60 | 2.01 | 7.12 | † | † | -0.10 | 1.59 | 4.45 | † | † | † | † |
| Swollen joint score | 0.65 | 2.66 | 8.06 | † | † | 0.15 | 1.90 | 5.25 | † | † | † | † |
| Tender joint count | 1.85 | 3.31 | 6.77 | † | † | -2.05 | 8.13 | 9.43 | † | † | † | † |
| Tender joint score | 3.60 | 6.85 | 10.47 | † | † | -4.65 | 16.94 | 11.82 | † | † | † | † |

*SD = standard deviation; SRM = standard response mean; ES = effect size; ESR = erythrocyte sedimentation rate; VAS = visual analog scale; BASFI = Bath Ankylosing Spondylitis Functional Index; SF-36 = Short Form 36. t-values unpaired (i.e. between group), t-values smaller than 2.03 are not statistically significant; between parentheses is the ranking order of the t-values. Minus (-) indicates deterioration. * due to a floor-effect responsiveness statistics and discriminatory capacity could not be calculated

Table 2.3 Indices of responsiveness at day 112 of follow up, for the treatment group and placebo group, ordered by the Guyatt method*.

| | Treatment group (n=20) Mean change | SD | SD baseline | SRM | ES | Mean change | SD | SD baseline | SRM | ES | Guyatt | t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ESR | 23.60 | 21.66 | 23.09 | 1.09 | 1.02 | -1.25 | 6.26 | 16.33 | -0.20 | -0.08 | 3.77 | 4.43 (3) |
| Physician global assessment | 30.45 | 20.24 | 19.62 | 1.50 | 1.55 | -1.20 | 12.61 | 16.43 | -0.10 | -0.07 | 2.41 | 5.94 (1) |
| VAS nocturnal pain | 2.55 | 1.93 | 2.06 | 1.32 | 1.24 | -0.20 | 1.22 | 2.45 | -0.16 | -0.08 | 2.10 | 5.40 (2) |
| Patient global assessment | 33.15 | 22.15 | 23.89 | 1.50 | 1.39 | 5.35 | 20.35 | 25.34 | 0.26 | 0.21 | 1.63 | 4.13 (4) |
| VAS overall pain | 1.05 | 0.94 | 0.73 | 1.12 | 1.44 | 0.35 | 0.67 | 0.69 | 0.52 | 0.51 | 1.57 | 2.70 (9) |
| BASFI | 10.93 | 10.38 | 8.20 | 1.05 | 1.33 | 3.22 | 7.72 | 10.09 | 0.42 | 0.32 | 1.42 | 2.67 (10) |
| Morning stiffness | 21.80 | 22.55 | 18.52 | 0.97 | 1.18 | 0.50 | 16.46 | 23.37 | 0.03 | 0.02 | 1.32 | 3.41 (6) |
| Chest expansion | 6.65 | 6.18 | 6.12 | 1.08 | 1.09 | 1.05 | 5.04 | 6.77 | 0.21 | 0.16 | 1.32 | 3.14 (7) |
| SF-36: role physical | 25.75 | 28.71 | 21.30 | 0.90 | 1.21 | 5.95 | 21.77 | 24.88 | 0.27 | 0.24 | 1.18 | 2.46 (13) |
| SF-36: bodily pain | 26.17 | 24.39 | 15.20 | 1.07 | 1.72 | 5.44 | 22.62 | 21.84 | 0.24 | 0.25 | 1.16 | 2.79 (8) |
| SF-36: vitality | 60.30 | 75.97 | 50.56 | 0.79 | 1.19 | 17.63 | 55.12 | 70.69 | 0.32 | 0.25 | 1.09 | 2.03 (17) |
| SF-36: physical functioning | 21.85 | 25.24 | 25.22 | 0.87 | 0.87 | -6.25 | 20.07 | 25.74 | -0.31 | -0.24 | 1.09 | 3.90 (5) |
| SF-36: physical component summary | 46.23 | 46.76 | 22.16 | 0.99 | 2.09 | 11.25 | 44.04 | 43.13 | 0.26 | 0.26 | 0.05 | 2.44 (14) |

| | Treatment group (n=20) Mean change | SD | SD baseline | SRM | ES | Mean change | SD | SD baseline | SRM | ES | Guyatt | t |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dougados functional index | 0.71 | 1.25 | 1.63 | 0.57 | 0.44 | 0.03 | 0.83 | 1.65 | 0.03 | 0.02 | 0.85 | 2.02 (18) |
| SF-36: social functioning | 12.98 | 17.34 | 18.83 | 0.75 | 0.69 | -0.25 | 16.10 | 16.75 | -0.02 | -0.01 | 0.81 | 2.50 (11) |
| SF-36: mental health | 0.67 | 1.30 | 1.55 | 0.52 | 0.43 | -0.22 | 0.93 | 1.52 | -0.24 | -0.14 | 0.72 | 2.50 (11) |
| SF-36: mental component summary | 9.20 | 16.73 | 19.88 | 0.55 | 0.46 | -2.20 | 14.71 | 15.71 | -0.15 | -0.14 | 0.63 | 2.29 (16) |
| SF-36: role emotional | 5.30 | 16.45 | 25.62 | 0.32 | 0.21 | 3.65 | 10.37 | 19.15 | 0.35 | 0.19 | 0.51 | 0.38 (21) |
| Modified Schober | 0.26 | 0.67 | 1.50 | 0.39 | 0.17 | -0.07 | 0.54 | 1.48 | -0.13 | -0.05 | 0.48 | 1.71 (20) |
| Fatigue Severity Scale | 3.89 | 10.07 | 12.97 | 0.39 | 0.30 | -3.16 | 9.24 | 9.24 | -0.34 | -0.34 | 0.42 | 2.30 (15) |
| SF-36: general health | 18.34 | 41.15 | 44.43 | 0.45 | 0.41 | -8.33 | 46.99 | 38.84 | -0.18 | -0.21 | 0.39 | 1.91 (19) |
| Modified enthesopathy index | 5.30 | 6.59 | 8.43 | † | † | 1.70 | 4.61 | 7.88 | † | † | † | † |
| Occiput-to-wall distance | 1.00 | 2.29 | 7.92 | † | † | -0.70 | 2.22 | 3.54 | † | † | † | † |
| Swollen joint count | 1.85 | 4.28 | 7.12 | † | † | -0.30 | 2.05 | 4.45 | † | † | † | † |
| Swollen joint score | 2.15 | 5.14 | 8.06 | † | † | -0.50 | 3.07 | 5.26 | † | † | † | † |
| Tender joint count | 4.10 | 5.76 | 6.77 | † | † | -1.25 | 6.70 | 9.43 | † | † | † | † |
| Tender joint score | 6.15 | 9.69 | 10.47 | † | † | -3.15 | 13.56 | 11.82 | † | † | † | † |

*SD = standard deviation; SRM = standard response mean; ES = effect size; ESR = erythrocyte sedimentation rate; VAS = visual analog scale; BASFI = Bath Ankylosing Spondylitis Functional Index; SF-36 = Short Form 36. t-values unpaired (i.e. between group), t-values smaller than 2.03 are not statistically significant; between parentheses is the ranking order of the t-values. Minus (-) indicates deterioration. * due to a floor-effect responsiveness statistics and discriminatory capacity could not be calculated

Table 2.4. Responsiveness statistics and unpaired t-values for the DC-ART ASAS core set for the treatment group

| | | Day 28 | | | | Day 112 | | | |
|---|---|---|---|---|---|---|---|---|---|
| Domain | Instrument | Guyatt | SRM | ES | t-value | Guyatt | SRM | ES | t-value |
| Function | BASFI | 1,31 | 1,12 | 0,79 | 4,85 | 2,10 | 1,32 | 1,24 | 5,40 |
| | Dougados Functional Index | 0,75 | 0,76 | 0,64 | 1,84 | 1,32 | 1,08 | 1,09 | 3,14 |
| Pain | VAS nocturnal | 1,66 | 1,25 | 1,35 | 5,02 | 1,63 | 1,50 | 1,39 | 4,13 |
| | VAS overall | 1,38 | 1,25 | 1,30 | 4,67 | 1,18 | 0,90 | 1,21 | 2,46 |
| Spinal mobility | Chest expansion | 0,99 | 0,76 | 0,37 | 2,08 | 0,85 | 0,57 | 0,44 | 2,02 |
| | Modified Schober | 0,40 | 0,30 | 0,13 | 0,78 | 0,48 | 0,39 | 0,17 | 1,71 |
| | Occiput to wall distance | 0,56 | 0,21 | 0,05 | 1,35 | 0,45 | 0,44 | 0,13 | 2,38 |
| Patient global | VAS last week | 1,64 | 1,27 | 1,44 | 3,43 | 1,57 | 1,12 | 1,44 | 2,70 |
| Stiffness | Morning Stiffness | 1,20 | 1,05 | 1,19 | 3,72 | 1,09 | 0,79 | 1,19 | 2,03 |
| Peripheral joints and entheses | Number of swollen joints | * | * | * | * | * | * | * | * |
| Acute phase reactants | ESR | 3,11 | 1,55 | 1,10 | 6,42 | 3,77 | 1,09 | 1,02 | 4,43 |

DC-ART = disease-controlling antirheumatic therapy; ASAS = Assessments in Ankylosing Spondylitis Working Group; SRM = standardized response mean; ES = effect size; t-values are unpaired (i.e. between group), t-values smaller than 2,03 are not statistically significant.

* due to a floor-effect responsiveness statistics and discriminatory capacity could not be calculated

# Discussion

For the domains pain and patient global assessment, this study has shown that the selected instruments have a large responsiveness and discriminatory capacity. While the measure morning stiffness in the domain stiffness has a good responsiveness, the t-value was barely statistically significant. This means that there is an effect in the treatment group [i.e. good responsiveness (Guyatt 1,09)], but that this effect creates a small, hardly significant contrast between the placebo and treatment group. Since the discriminative capacity is influenced by the number of patients, it is possible that the measure will have better discriminatory capacity when evaluated in a larger number of patients.

For the domain spinal mobility, only chest expansion demonstrated a good responsiveness but the discriminatory capacity was not very good. Again, this may be explained by the small number of patients. It is necessary to recognize that the selected patient population is likely an important factor in the assessment of spinal mobility. There is a large chance that, given the mean disease duration of 13 years for patients in this trial, a number of the patients may have had some degree of spinal ossification, limiting their ability to improve in some of these outcomes. It must also be taken into account that for this domain the period of follow-up might have been too short. Further work is needed to establish for which patients the measures of spinal mobility might be responsive (e.g. early disease) and which period of follow-up is necessary (e.g. ≥1 year).

For the domain peripheral joints and entheses only the number of swollen joints was selected by ASAS. In this study this measure was vulnerable to a floor effect and therefore no conclusions can be made about discriminatory capacity and responsiveness. The modified enthesopathy index was not selected by ASAS but also was unable to be evaluated because of a floor effect. The last domain of the ASAS DC-ART core set, the measure ESR proved to be the most responsive and discriminative in this trial. It must be kept in mind that the results need to be interpreted in the context of this trial; this means that it concerns a selected study population of patients with a high disease activity in established disease, treated with TNF blocking therapy.

Concerning other indices and measures studied in this trial several issues can be noted. The measures for assessment of the peripheral joints and entheses all suffer from a floor effect: a large percentage of patients have no affected joints and/or entheses. These measures are context specific, and their appropriateness will depend upon the population being investigated. The physician's global assessment measured by means of a VAS has an excellent responsiveness and discriminative capacity. From this point of view it may be an asset to the core set. The question is if the measure has an additional intrinsic value, as it is known that AS patients can accurately assess their own disease activity[24]. Furthermore, it is also likely that physicians take into account other

variables such as ESR when assessing the physician global, which consequently makes the physician global assessment a dependent measure.

The ASAS working group did not select an instrument for fatigue, as not enough data on the performance of variably instruments in AS were available. The one used in this study, the Fatigue Severity Scale demonstrated a moderate responsiveness and good discriminative capacity. In addition, the vitality scale on the SF-36 also demonstrated good responsiveness and discriminatory capacity and with respect to content it is fairly comparable with the Fatigue Severity Scale. Therefore, this individual scale may be able to be used as a means of measuring fatigue.

The final measure was the SF-36. It turned out that the different scales of the SF-36 had varying degrees of responsiveness and discriminative power. It seems that the items focusing on physical function and pain were more responsive than the items dealing with mental health and emotions. For patients with AS, treatment with a TNF blocking agent seems to have more effect on pain and physical functioning than on mental health and emotions. This is coherent with the earlier described fact that after pain and stiffness, one of the most important complaints of patients with AS is disability[25]. These aspects are already assessed by the other measures in the ASAS core set and since the scales related to mental health and emotions demonstrate a small responsiveness, the SF-36 appears to contribute little in the setting of a clinical trial. However, it is possible that the SF-36 could be helpful in allowing for comparison of quality of life of patients with other diseases.

So far the ASAS working group has only chosen single measures. However, the value of combined indices such as the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI)[26] and the Bath Ankylosing Spondylitis Metrology Index (BASMI)[27] may warrant further investigations. Considering responsiveness, it could be expect that indices are more responsive than single measures, as combining reduces scatter.

Improvement/response criteria have already been developed in several areas within rheumatology, and recent improvement criteria for AS has been studied by Anderson et al.[28] They developed criteria for improvement in AS based upon the five domains of the ASAS SMARD core set (function, pain, spinal mobility, patient global and stiffness) by using outcome data from placebo-controlled clinical trials of NSAIDs. It was concluded that all the domains were appropriate except spinal mobility, because of a lack of responsiveness of the mobility measures. Results from our study support their conclusions, except for the domain spinal mobility. This domain has three measures, chest expansion, modified Schober and occiput-to-wall distance. While it was not possible to do responsiveness analysis for the occiput-to-wall distance, we did demonstrate an acceptable responsiveness for chest expansion. However, the discriminative capacity was relatively low. An explanation for the finding that in NSAIDs trials the mobility measures are not responsive and in our study chest expansion is responsive, might be that it is less likely that spinal mobility is impacted by NSAID treatment than by TNF blocking therapy.

In this study we have shown a tight relationship between responsiveness and discriminatory capacity. Most measures with good responsiveness also showed good discriminatory capacity, and vice versa. In this particular case, this tight relationship is due to the large contrast between active intervention and placebo, and partly to the definition of the responsiveness statistic we have chosen, since the Guyatt effect size encompasses the variation observed in the placebo-group. A responsive measurement will not be discriminative in all situations, however. Responsiveness statistics are considered measurement-specific, and can be used across different studies; discrimination statistics depend on the responsiveness of a measurement plus context-specific factors, such as sample size, treatment contrast, variation in the control group, etcetera.

Differences in responsiveness, and especially discrimination, may have important implications for clinical trial design. The use of measures that are both responsive and discriminative increase the statistical power of a clinical trial.

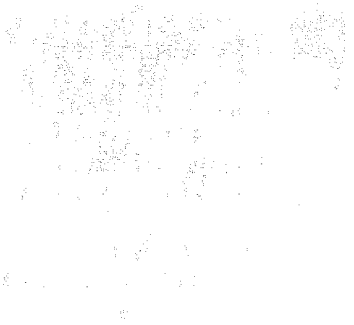
It is for the first time that the responsiveness and discriminatory capacity of the ASAS DC-ART core set have been evaluated. The sample size of this study may limit the generalisability, since only a small selection of patients with active longstanding disease have been investigated. On the other hand, finding these results in such a severely afflicted patient population adds to the validity of the results. It is important to realize that etanercept may specifically influence certain measures, and that other therapies may lead to different effects. Therefore, responsiveness and discriminatory validity should also be assessed in trials with other treatments.

In summary, this study has confirmed responsiveness and discriminatory capacity of all measures included in the ASAS-DC-ART core set, with the exception of the domains spinal mobility (although the instrument chest expansion is responsive) and peripheral joints. In addition to measures of the ASAS DC-ART core set, other measures have shown to be very responsive and discriminative. The most important were physician's global assessment of disease activity, and the physical functioning- and pain scales of the SF-36.

# References


1. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. Assessments in Ankylosing Spondylitis Working Group. J Rheumatol 1997;24:2225-9.
2. van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. Assessments in Ankylosing Spondylitis. J Rheumatol 1999;26:951-4.
3. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol 1998;25:198-9.
4. Baeten D, Kruithof E, Van den Bosch F, Demetter P, Van Damme N, Cuvelier C, De Vos M, Mielants H, Veys EM, De Keyser F. Immunomodulatory effects of anti-tumor necrosis factor alpha therapy on synovium in spondylarthropathy: histologic findings in eight patients from an open-label pilot study. Arthritis Rheum 2001;44:186-95.
5. Brandt J, Haibel H, Cornely D, Golder W, Gozalez J, Reddig J, Thriene W, Sieper J, Braun J. Successful treatment of active ankylosing spondylitis with the anti-tumor necrosis factor alpha monoclonal antibody infliximab. Arthritis Rheum 2000;43:1346-52.
6. Gorman JD, Sack KE, Davis JC. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor alpha. N Engl J Med 2002;346:1349-56
7. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27: 361-8.
8. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, Jenkinson T. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281-5.
9. Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988; 15:302-7.
10. Moll JM, Wright V. An objective clinical study of chest expansion. Ann Rheum Dis 1972;31(1):1-8.
11. Cash JM. Evaluation of the patient: history and physical examination. In: Klippel JH, editor. Primer on the Rheumatic diseases. Atlanta: Arthritis Foundation; 1997;92.
12. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989;46: 1121-3.
13. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
14. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12(4 S): 142S-158S.
15. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40:171-8.
16. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000;53 459-68.
17. Ruof J, Sangha O, Stucki G. Comparative responsiveness of 3 functional indices in ankylosing spondylitis. J Rheumatol 1999;26:1959-63.
18. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997; 50:869-79.

19. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369-78.
20. Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press 1977.
21. Ruof J, Stucki G. Comparison of the Dougados Functional Index and the Bath Ankylosing Spondylitis Functional Index. A literature review. J Rheumatol 1999;26:955-60.
22. Daltroy LH, Larson MG, Roberts NW, Liang MH. A modification of the Health Assessment Questionnaire for the spondyloarthropathies. J Rheumatol 1990;17:946-50.
23. Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, De Vlam K, Mielants H, van der Tempel H, van der Linden S. A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index and the Dougados Functional Index in the assessment of ankylosing spondylitis. J Rheumatol 1999;26:961-5.
24. Hidding A, van Santen M, De Klerk E, Gielen X, Boers M, Geenen R, Vlaeyen J, Kester A, van der Linden S. Comparison between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. J Rheumatol 1994;21:818-23.
25. Calin A. The individual with ankylosing spondylitis: defining disease status and the impact of the illness. Br J Rheumatol 1995;34:663-72.
26. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994;21:2286-91.
27. Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. Ankylosing spondylitis: the correlation between a new metrology score and radiology. Br J Rheumatol 1995;34:767-70.
28. Anderson JJ, Baron G, van der Heijde D, Felson DT, Dougados M. Ankylosing spondylitis assessment group preliminary definition of short-term improvement in ankylosing spondylitis. Arthritis Rheum 2001;44:1876-86.



# Chapter 3

# What is the most appropriate radiologic scoring method for ankylosing spondylitis? A comparison of the available methods based on the OMERACT filter

A Wanders, R Landewé, A Spoorenberg, M Dougados, Sj van der Linden, H Mielants, H van der Tempel, D van der Heijde

*Arthritis Rheum 2004; 50:2622-32*

# Introduction

For assessing the disease modifying potential of drugs in Ankylosing Spondylitis (AS), the demonstration of reduction or stopping of structural damage is essential. Structural damage in AS can be measured on radiographs of the spine and hips. A number of radiographic scoring methods are available for this purpose: The Bath Ankylosing Spondylitis Radiology Index (BASRI)[1], the Stoke Ankylosing Spondylitis Spine Score (SASSS)[2] and a modification of the SASSS, the so-called modified SASSS[3]. The BASRI exists in 2 forms: the BASRI-spine and the BASRI-total. The former excludes, and the latter includes the hips. The BASRI and SASSS have been published in the literature, and the modified SASSS was published in a thesis only. All methods have been validated by their developers.

It is commendable that one of these methods would be selected as the radiographic outcome assessment of choice for clinical trials, in order to assure uniformity and allow a comparison of data across trials in the future. The ASsessment in Ankylosing Spondylitis (ASAS) working group works on the standardization of measurements in AS[4], and the selection of a method to assess radiographic progression is one of the important issues on her research agenda.

The validity of radiographic scoring methods in AS is hardly investigated in the past. Spoorenberg et al. from our group has initiated a method-comparison with a maximum follow-up of two years, primarily relating to assessing damage scores[5]. In this study, some aspects of reliability (intra- and inter observer reliability of status scores) of all three methods were established. Moreover, agreement between two observers on progression in individual patients was assessed, but only with a strict definition on agreement. In clinical trials, however, the subject of interest is change of radiographic damage, primarily on the group level, and not the absolute level of damage itself. Apart from that, and according to OMERACT, discrimination (sensitivity-to-change), truth (construct validity) and feasibility of scoring methods should be investigated before a preference can be made[7].

The main objective of the present study was therefore to test the radiographic scoring methods for all three aspects of the OMERACT-filter over a follow-up period of four years, including an evaluation of reliability of progression scores.

# Patients and methods

## Patients and films

Radiographs from the OASIS cohort, an international longitudinal, observational study on outcome in AS were used[5]. Originally 217 patients from four centers in the Netherlands, Belgium and France were included in this cohort. Radiographs were obtained at baseline, and after respectively one, two and four years of follow-up. One set of radiographs consisted of a posterior-anterior view of the pelvis, to score the SI

joints and the hips, an anterior-posterior and lateral view of the lumbar spine, and a lateral view of the cervical spine. Due to loss to follow-up, sets of radiographs including the 4-year film, of 133 patients were available; only these patients were included in this study.

### The scoring methods

Currently there are two scoring methods widely known, the BASRI and SASSS.
In 1995 Kennedy et al.[8] used a radiology score, which was the precursor of the BASRI. In this score the sacroiliac joints (SI joints) were graded according to the New York criteria[9], these criteria describe five grades of sacroiliitis ranging from 0 to 4 (table 3.1). For the hips, cervical and lumbar spine a comparable system was developed. In 1998 Mackay et al.[1] published the currently used BASRI, the BASRI-spine, which includes the SI joints (still scored according to the New York criteria) and the lumbar and cervical spine. To assess the lumbar spine, a lateral and anteroposterior (AP) view are used. The score for the lumbar spine is a composite score of both views: E.g. if one view shows syndesmophytes at one particular level, and the other view shows syndesmophytes at a different level, syndesmophytes are considered present at two different levels, and scored accordingly The lumbar spine was defined as extending from the lower border of T12 to the upper border of S1. To assess the cervical spine a lateral view was used and the cervical spine was defined as extending from the lower border of C1 to the upper border of C7. The lumbar spine and cervical spine are separately graded from 0 to 4 (table 3.1). The BASRI-spine is the sum of the mean score of the right and left sacroiliac joint (in one decimal) plus the score of the lumbar spine plus the score of the cervical spine. According to the NY criteria AS patients are supposed to have radiographic sacroiliitis, so the range for the BASRI-spine in patients fulfilling the criteria is 2 – 12. In 2000, the group from Bath published another paper[10] in which they introduced the BASRI-total, which adds the BASRI-hip to the BASRI-spine. The score for the hip is based on the same grading system as that applies to the other parts of the BASRI (table 3.1). The BASRI hip score is the mean of the right and left hip (with one decimal). In the publications concerning the BASRI it is not explained how to handle missing observations. It was decided that when it was impossible to assess 75% or more of a view or if a view was missing, the patient was excluded from our study.
The second method that is evaluated in this study is the SASSS. In 1991 Taylor et al.[11] published this method, which is a detailed scoring system for the anterior and posterior site of the lumbar spine, with a range from 0 to 72. The SASSS is obtained by viewing the lower border of the 12$^{th}$ thoracic vertebra, all five lumbar vertebrae and the upper border of the sacrum on a lateral view. All four corners of each vertebra are examined and scored 1 for an erosion, sclerosis and / or squaring, 2 for a syndesmophyte and 3 for total bony bridging at each site, giving a maximum possible score of 72 (table 3.1). In a publication of Averns et al.[2] in 1996 the name Stoke Ankylosing Spondylitis Spine Score (SASSS) was introduced. In both publications concerning the SASSS it is not explained how to handle missing observations. We decided that if more than three

scoring sites were missing, radiographs were not taken into account. If three or less sites were missing the mean of the other scoring sites was used as a substitute for the missing sites.

Table 3.1 Radiological scoring methods

| | | |
|---|---|---|
| **New York criteria for sacroiliitis** | | |
| (mean score of both SI joints is used in BASRI) | | |
| *Score* | *Grade* | |
| 0 | Normal | |
| 1 | Suspicious | No definite change |
| 2 | Minimal | Minimal sacroiliitis defined by loss of definition at the edge of the sacroiliac joints, some juxta-articular sclerosis, minimal erosions, there may be some narrowing |
| 3 | Moderate | Moderate sacroiliitis defined by definite sclerosis on both sides, blurring and indistinct margins and erosive changes with loss of joint space |
| 4 | Severe | Complete fusion or ankylosis of the joints |
| **Bath Ankylosing Spondylitis Radiology Index (BASRI) for the hips** | | |
| (mean score of both hips is used in BASRI-total) | | |
| 0 | Normal* | No change |
| 1 | Suspicious* | Focal joint space narrowing |
| 2 | Mild* | Circumferential joint space narrowing >2 mm |
| 3 | Moderate* | Circumferential joint space narrowing ≤2 mm or bone-on-bone apposition of <2 cm |
| 4 | Severe* | Bone deformity or bone-on-bone apposition ≥2 cm |
| **Bath Ankylosing Spondylitis Radiology Index (BASRI) for the spine** | | |
| (Lumbar spine: AP and lateral view of lumbar spine are scored, the view with the highest score is taken. Cervical spine: the lateral view is scored) | | |
| 0 | Normal | No change |
| 1 | Suspicious | No definite change |
| 2 | Mild | Any number of erosions, squaring, or sclerosis, with or without syndesmophytes, on ≤2 vertebrae |
| 3 | Moderate | Syndesmophytes on ≥3 vertebrae, with or without fusion involving 2 vertebr |
| 4 | Severe | Fusion involving ≥3 vertebrae |
| **Stoke Ankylosing Spondylitis Spine Score (SASSS) and modified SASSS (range 0-72).** | | |
| For the SASSS the anterior and posterior site from the lower border of the 12$^{th}$ thoracic vertebra up to the upper border of the first sacral vertebra are scored. For the modified SASSS only the anterior site of the lumbar spine and the anterior site of the cervical spine from the lower border of the second cervical vertebra up to the upper border of the thoracic vertebra. | | |
| 0 | Normal | |
| 1 | Erosion, sclerosis or squaring | |
| 2 | Syndesmophyte | |
| 3 | Bridging syndesmophyte | |

*The grade should be increased by 1 if 2 out of the following bony changes are present: erosions, osteophytes, and protrusion.

The final method that we included in this study is a method derived from a publication by Creemers in her thesis[3]. This method is a modification of the SASSS and scores the anterior site of the lumbar and cervical spine at a lateral view. The anterior site of the same vertebrae of the lumbar spine as described for the SASSS, are scored, and the anterior site of the cervical spine from the lower border of the 2nd cervical vertebra up to the upper border of the first thoracic vertebra. So the range remains 0–72. We dealt with missing observations similarly as described for the SASSS.
One observer (AW) scored the films according to the different methods. For each patient, the order in which the methods were applied was always the same. First the SI-joints and hips were scored according to the BASRI. Secondly the lumbar spine was scored according to the BASRI and the SASSS. Finally the cervical spine was scored according to the BASRI and modified SASSS. During the follow up the format of the radiographs was changed in some centers, which made it possible for the observer to identify the point in time. Therefore all films were scored with known chronology for all methods.

### The OMERACT filter as instrument to evaluate the different scoring methods

All methods: BASRI (split in BASRI-spine and BASRI-total), SASSS and the modified SASSS were judged with respect to the different aspects of the OMERACT filter: truth, discrimination and feasibility[7].

#### *Truth*

The aspect truth deals with the following questions: Is the measure truthful, does it measure what it is intended? How is the validity of the measure? To establish a valid radiological scoring method for AS in clinical trials, it is important that the method includes the relevant parts of the skeleton (construct validity). Therefore we evaluated every single part of the skeleton included by the different scoring methods. By doing this we got an impression about the involvement of the different parts, and we found out in which parts changes occurred. For the lumbar spine, we also compared the additional information obtained from the anteroposterior view with the additional information obtained from the lateral view, since both views are used in the BASRI and only the lateral view in the SASSS and modified SASSS. Also the anterior spine versus the posterior spine was evaluated, since in the SASSS both sites are scored and in the modified SASSS only the anterior site. The construct validity of the method was also assessed by examining the correlation of the scoring methods with measures of spinal mobility, disease duration and functional limitation. As measures for spinal mobility the occiput-to-wall distance, the modified Schober and lateral spinal flexion[12] were used. Lateral spinal flexion is the difference between the distance between the patient's middle fingertip and the floor while the patient is standing and while the patient has bend side wards maximally. As measure for functional limitation the Bath Ankylosing Spondylitis Functional Index was used[13]. Correlation was expressed as Spearman's rho.

*Discrimination*

The aspect discrimination concentrates on the question; does the measure discriminate between situations of interest. The word captures issues of reliability and sensitivity to change. To assess interobserver reliability, another observer (RL) scored sets of radiographs of 20 patients with four time points per patient with known chronological order. These same 80 sets were scored in chronological order with a time interval of 4 weeks by the first observer (AW) to assess intraobserver reliability. So interobserver and intraobserver reliability of the different scoring methods were assessed on 80 status scores and 60 progression intervals.
Inter- and intraobserver reliability can be expressed as intraclass correlation coefficients (ICC's) or as variance components calculated by ANOVA. We have chosen an ANOVA analysis with the variance components patients, observer and residual, because comparison of variance components provide better insight in the kind of error. Each variance component represents the percentage of the total variance that can be explained by that particular component. The variance component patient reflects the variance that is caused by true differences among patients, the variance component observer reflects the variance attributable to differences between observers and the variance component residual reflects the remaining random error. The variance component of patients can be used to compare with studies in which ICC's are presented, since it is equal to the ICC if the latter one is calculated with observer as fixed factor.
To obtain insight in the sensitivity to change of the methods, the means and medians are given for baseline and after one, two and four years of follow-up. Also the percentages of patients that show changes greater than zero are given and effect sizes were calculated. Effect sizes were calculated on logarithmically transformed data because of the skewed distribution pattern.

*Feasibility*

This last aspect of the OMERACT filter concentrates on the question whether the measure can be applied easily, given constraint of time, money and interpretability. In order to give insight in these matters we provide information about time needed for training, scoring of the methods and radiation exposure for the patients.

# Results

Patients and study course

In table 3.2 the characteristics of the patients of the OASIS cohort included in this study are described as well as the characteristics of the patients not included. It appeared that the patients that were not included were younger and had shorter disease duration, but that they were affected somewhat more severely than the included group. Although radiological damage, disease activity, function and spinal mobility measures were

somewhat worse for the not included group, the differences were not statistically significant. Therefore the included population is considered representative for the entire OASIS cohort.

### Handling of missing data

The sites that were most scored as missing at the lateral view of the cervical spine, were the lower three; the upper and lower border of the 7th cervical vertebra (respectively in 4% and 10% of patients) and the upper border of the first thoracic vertebra (in 10% of patients). One way of dealing with this problem would be to exclude these sites of the scoring system, which would lead to a loss of 25% of information regarding the cervical spine. Therefore, we felt substitution in about 10% of patients is a preferable approach.

### The OMERACT filter

#### *Truth*

The involvement of - and progression in - the different parts of the skeleton are presented in table 3.3. For the SI joints and hips it can be seen that only a small percentage of patients shows progression, and that the involvement of the hips is limited Abnormalities in the spine are scored for the majority of patients at baseline with all methods, and also changes can be scored by all methods in the follow up of four years. The same is observed for the cervical spine. When the anterior and posterior sites of the lumbar spine are observed separately, it can be seen that the majority of patients show damage at the anterior site, and that this site also shows most progression. Noteworthy is the difference in the percentage of patients that show progression for the lumbar and cervical spine using the different methods. The SASSS and modified SASSS quantify a higher percentage of patients as being progressive than the BASRI does.

We compared the radiological damage and progression visible on the AP view of the lumbar spine with the lateral view. Both views do not provide the same information. In 12% of all cases more damage was seen on an AP view. So if the AP view is omitted for staging AS patients, valuable information of 12% of patients would be missed.

We also investigated whether loss of information was similar if progression scores were used. In half of the cases in which the progression on both views differs, the progression scored on an AP view is greater. Whether this progression on the AP view is significantly greater than on the lateral view, was investigated by focusing on these sets of radiographs. Of all 389 intervals studied, it turned out that in 39 (10%) the AP view showed more progression than the lateral view. These intervals concerned 19 patients. For each patient the progression on the AP view was compared with the progression on the lateral view. Only in 4 patients (3%) the missed information on the AP view would have added information to the scoring derived from only the lateral view. So for the aim of staging, the AP lumbar view provides relevant additional information, but if films are assessed with the aim of evaluating progression, the AP view does not importantly contribute.

Table 3.2 Patient characteristics of the OASIS cohort at baseline

| | Patient included in this study (n=133) | | | | | | Patients not included in this study (n=84) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variable | mean | SD | median | P25 | P75 | range | mean | SD | median | P25 | P75 | range |
| Age (years) | 44.6 | 11.7 | 44.0 | 35.0 | 53.3 | 20.3-78.0 | 42.1 | 14.0 | 40.2 | 31.6 | 50.3 | 19.0-77.0 |
| Mean duration of complaints (years) | 21.0 | 11.6 | 18.2 | 12.3 | 27.1 | 0.4-51.0 | 18.9 | 12.0 | 15.9 | 9.6 | 25.5 | 0.0-54.0 |
| Mean duration of disease after diagnosis (years) | 11.7 | 9.3 | 10.0 | 4.5 | 15.7 | 0.2-42.0 | 10.7 | 9.5 | 9.5 | 4.2 | 9.5 | 0.0-34.0 |
| Male (%) | 69.7 | | | | | | 73.5 | | | | | |
| Radiological outcome | | | | | | | | | | | | |
| BASRI-spine score (0-12) | 6.5 | 3.0 | 7,0 | 4.0 | 9.0 | 1.0-12.0 | 6.6 | 3.7 | 5.5 | 3.0 | 10.0 | 0.0-12.0 |
| BASRI-total score (0-16) | 6.9 | 3.4 | 7.0 | 4.0 | 9.0 | 1,0-16.0 | 7.0 | 4.5 | 6.5 | 3.0 | 10.0 | 0.0-16.0 |
| SASSS score (0-72) | 10.1 | 18.0 | 2.0 | 0.0 | 12.0 | 0.0-72.0 | 12.6 | 20.5 | 2.0 | 0.0 | 17.0 | 0.0-72.0 |
| Modified SASSS score (0-72) | 12.7 | 17.4 | 5.0 | 0.0 | 16.9 | 0.0-72.0 | 16.5 | 22.9 | 4.0 | 0.0 | 31.3 | 0.0-72.0 |
| Disease activity | | | | | | | | | | | | |
| BASDAI (0-10) | 3.4 | 2.1 | 3.2 | 1.7 | 5.0 | 0.0-8.5 | 3.5 | 2.2 | 3.2 | 1.7 | 4.9 | 0.0-9.7 |
| Function | | | | | | | | | | | | |
| BASFI (0-10) | 3.3 | 2.5 | 3.2 | 1.0 | 5.0 | 0.0-9.7 | 3.5 | 2.7 | 3.2 | 1.0 | 5.3 | 0.0-10.0 |
| Spinal mobility | | | | | | | | | | | | |
| Lateral spinal flexion (cm) | 10.2 | 6.4 | 10,0 | 5.5 | 15.1 | 0.0-24.2 | 9.7 | 6.3 | 9.4 | 4.5 | 14.4 | 0.0-26.1 |
| Occiput-to-wall distance (cm) | 3.6 | 4.9 | 1.5 | 0.0 | 6.2 | 0.0-26.0 | 4.3 | 6.6 | 0.0 | 0.0 | 5.8 | 0.0-26.1 |
| Modified Schober (cm) | 12.9 | 2.0 | 13.2 | 11.9 | 14.1 | 1.4-16.8 | 12.5 | 1.5 | 12.5 | 11.0 | 13.7 | 10.0-16.0 |

OASIS = Outcome in Ankylosing Spondylitis International Study; BASRI = Bath Ankylosing Spondylitis Radiology Index; SASSS = Stoke Ankylosing Spondylitis Spinal Score; BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; BASFI = Bath Ankylosing, Spondylitis Functional Index; SD = standard deviation; p25 = 25$^{th}$ percentile; p75 = 75$^{th}$ percentile. For all ranges it applies that the lowest value indicates the best situation.

Table 3.3 Percentages of patients with structural damage at baseline and radiographic progression in the follow-up of four years, for the different parts of the skeleton, scored according to the three methods.*

| | BASRI | | SASSS | | Modified SASSS | |
|---|---|---|---|---|---|---|
| | % patients with baseline damage | % patients that show changes | % patients with baseline damage | % patients that show changes | % patients with baseline damage | % patients that show changes |
| SI joints | 100 | 9 | NA | NA | NA | NA |
| Hips | 24 | 8 | NA | NA | NA | NA |
| Lumbar spine | 68 | 18 | 60 | 46 | 60 | 43 |
| Anterior site | NA | NA | 60 | 43 | 60 | 43 |
| Posterior site | NA | NA | 18 | 15 | NA | NA |
| Cervical spine | 65 | 23 | NA | NA | 56 | 41 |
| Anterior site | NA | NA | NA | NA | 56 | 41 |

* Change is defined as every change greater than zero. BASRI = Bath Ankylosing Spondylitis Radiology Index; SASSS = Stoke Ankylosing Spondylitis Spine Score; SI = sacro iliac; NA = not applicable.

We also compared the radiological damage and progression visible on the anterior site of the lumbar spine with the posterior site. In more than half of the patients there is a difference in damage of the anterior site versus the posterior site of the spine. Almost anytime this difference is caused by the fact that the damage at the anterior site is worse. But 13 patients showed more progression at the posterior site than at the anterior site. For each patient the progression at the anterior site was compared with the progression at the posterior site. We found that in 10 of these 13 patients the progression at the posterior site significantly contributes to the total progression. For staging purposes, assessing the posterior site does not contribute, but for assessing progression it contributes significantly in less than 10% of the patients.
In table 3.4 the correlation between measurements of spinal mobility, disease duration and BASFI and the different scoring methods is given. The table indicates significant correlations between spinal mobility, disease duration, BASFI and radiological damage. For all methods the correlations show the same magnitude.

### *Discrimination*

The first part of the aspect discrimination is reliability. Inter- and intraobserver reliability is given in table 3.5 for status scores at 2 years and for progression scores for a follow-up of 2 years. The interobserver reliability for the status scores is for all methods very good and for the modified SASSS excellent. The intraobserver reliability for status scores is excellent for all methods. The interobserver reliability for progression scores shows only a good reliability for the modified SASSS, reliability for the BASRI and SASSS is unsatisfactory. When attention is paid to the kind of error for the different methods it appears that for BASRI the error is random and that for the SASSS the error

consists of random error and error that is caused by differences between the observers. It is remarkable that there is such a difference between the SASSS and modified SASSS. Therefore the different sites of the SASSS were investigated related to interobserver reliability. It appears that for the anterior site the variance component residual is 32.4, the variance component observer is 6.8 and the variance component patient is 60.9. For the posterior site these numbers are respectively 71.4, 25.9 and 2.7. So the interobserver reliability of mainly the posterior site is absolutely poor. The intraobserver reliability of progression scores was good for all methods. Only reliability scores at 2 years and after an interval of 2 years are given but all results applies also for the status scores at baseline, 1 year and 4 year and for the progression scores after a 1- and 4-year interval (data not shown). There is one exception and that is the interobserver reliability of the progression score according to the modified SASSS after a one-year interval. This score was not good; variance component patient = 49.6. This poor interobserver reliability was due to residual error (40.7).
Data concerning sensitivity to change is presented in table 3.6. In table 6 it can be seen that changes can be detected by all methods. After four years of follow-up the BASRI-spine and BASRI-total show a change of 0.6 and 0.7 points, and the SASSS and modified SASSS 3.5 and 4.4. In table 3.7 the percentages of patients that show changes for each method are represented. The modified SASSS quantifies the highest percentage of patients with changes. We further investigated if a ceiling effect occurred. At baseline 5.3% of patients had a maximal score for BASRI-spine, this percentage was 0.8 for BASRI-total, and 3.8% and 0.8% for the SASSS and modified SASSS respectively. If the different parts of the scoring system were considered separately, then 14% of patients had a maximal score for the lumbar spine according to BASRI versus 5 % for SASSS and modified SASSS and for the cervical spine 12% of patients had a maximal score according to the BASRI versus 2% of patients for the modified SASSS. Therefore progression scores as assessed by BASRI may be affected by a ceiling effect. The ranges of effect sizes for the different intervals of the methods are as follows: BASRI-spine and BASRI-total have lower effect sizes (0.12–0.36) than the SASSS (0.32–0.51) and mSASSS (0.34–0.58).

### *Feasibility*

The time needed for each scoring methods differs. It is not possible to give a mean required time for scoring a single set for a patient. It can be noticed that the BASRI will take the least time since this method is less detailed than the SASSS and modified SASSS as with these last two methods every single corner of a vertebra needs to be assessed. The same applies for the time needed for training. The radiation exposure for the patients is as follows (based on data provided by the radiology department of the University Hospital Maastricht): AP view of the pelvis = 0.54 mSv, AP view of the lumbar spine = 0.54 mSv, lateral view of the lumbar spine = 0.93 mSv and a lateral view of the cervical spine = 0.07 mSv. The total exposure for the different methods is 2.08 mSv for the BASRI, for the SASSS 0.93 mSv and for the modified SASS 1 mSv.

Table 3.4 Range of correlations between radiological damage and spinal mobility, disease duration and BASFI expressed as Spearman's rho calculated at baseline, 1 year, 2 years and 4 years of follow-up.

| | BASRI-spine | BASRI-total | SASSS | modified SASSS |
|---|---|---|---|---|
| Lateral spinal flexion | -0.50 – -0.75 | -0.47 – -0.75 | -0.56 – -0.77 | -0.52 – -0.75 |
| Occiput-to-wall distance | 0.59 – 0.65 | 0.56 – 0.63 | 0.53 – 0.61 | 0.52 – 0.64 |
| Modified Schober | 0.51 – -0.65 | 0.50 – -0.65 | -0.61 – -0.76 | -0.56 – -0.67 |
| Disease duration | 0.37 – 0.42 | 0.38 – 0.42 | 0.34 – 0.36 | 0.33 – 0.36 |
| BASFI | 0.33 – 0.39 | 0.34 – 0.39 | 0.33 – 0.41 | 0.32 – 0.37 |

BASFI = Bath Ankylosing Spondylitis Radiology Index; BASRI = Bath Ankylosing Spondylitis Radiology Index; SASSS = Stoke Ankylosing Spondylitis Spine Score

Table 3.5 Inter- and intraobserver reliability based on 20 patients expressed in variance components (% of total variance)

| | Interobserver reliability | | | | | | Intraobserver reliability | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Status score at 2 years | | | Progression score after 2 years | | | Status score at 2 years | | | Progression score after 2 years | | |
| | vc res* | vc obs | vc pat | vc res | vc obs | vc pat | vc res | vc obs | vc pat | vc res | vc obs | vc pat |
| BASRI-total | 13.4 | 0.0 | 86.6 | 50.9 | 0.6 | 48.5 | 3.4 | 0.4 | 96.3 | 6.3 | 0.7 | 93.0 |
| BASRI spine | 14.9 | 0.0 | 85.1 | 48.9 | 0.0 | 51.1 | 2.7 | 0.3 | 97.0 | 6.3 | 0.0 | 93.0 |
| SASSS | 7.4 | 4.0 | 88.6 | 32.5 | 23.8 | 43.7 | 0.9 | 0.0 | 99.1 | 21.3 | 0.0 | 78.7 |
| Modified SASSS | 2.2 | 0.0 | 98.4 | 17.9 | 0.4 | 81.7 | 0.8 | 0.2 | 99.1 | 5.0 | 0.0 | 95.0 |

vc res = variance component residual; vc obs = variance component observer; vc pat = variance component patient, BASRI = Bath Ankylosing Spondylitis Radiology Index; SASSS = Stoke Ankylosing Spondylitis Spine Score. * Each variance component represents the percentage of the total variance that can be explained by that particular component.

Table 3.6 Descriptive statistics of four year follow-up of structural damage according to the different scoring methods in 133 Ankylosing Spondylitis patients.

| | BASRI-spine | | | | | BASRI-total | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | n | mean | SD | median | P25 – p75 | n | mean | SD | median | P25 – p75 |
| Baseline | 133 | 6.5 | 3.0 | 7.0 | 4.0 – 9.0 | 133 | 6.9 | 3.4 | 7.0 | 4.0 – 9.0 |
| 1 year | 129 | 6.7 | 3.1 | 7.0 | 4.0 – 9.0 | 129 | 7.1 | 3.4 | 7.5 | 4.0 – 9.0 |
| 2 years | 127 | 6.9 | 3.0 | 7.0 | 4.0 – 9.0 | 127 | 7.3 | 3.4 | 8.0 | 4.5 – 10.0 |
| 4 years | 133 | 7.1 | 3.1 | 7.5 | 4.5 – 9.8 | 133 | 7.6 | 3.5 | 8.0 | 4.8 – 10.0 |
| | SASSS | | | | | Modified SASSS | | | | |
| | n | mean | SD | median | P25 – p75 | n | mean | SD | median | P25 – p75 |
| Baseline | 132 | 10.1 | 18.0 | 2.0 | 0.0 – 12.0 | 131 | 12.7 | 17.4 | 5.0 | 0.0 – 16.9 |
| 1 year | 129 | 11.7 | 18.8 | 3.0 | 0.0 – 14.0 | 128 | 14.4 | 18.3 | 5.5 | 0.0 – 20.0 |
| 2 years | 127 | 12.6 | 19.2 | 4.0 | 0.0 – 16.0 | 126 | 15.5 | 18.9 | 6.5 | 0.0 – 22.4 |
| 4 years | 133 | 13.6 | 19.3 | 6.0 | 0.0 – 20.0 | 132 | 17.1 | 19.6 | 9.9 | 1.0 – 27.6 |

SD = standard deviation; p25 – p75 = 25$^{th}$ and 75$^{th}$ percentile; BASRI = Bath Ankylosing Spondylitis Radiology Index; SASSS = Stoke Ankylosing Spondylitis Spinal Score.

Table 3.7 Proportion of patients that show a change ≥ 1 unit per scoring method

| | 1 year interval | 2-year interval | 4-year interval |
|---|---|---|---|
| BASRI-spine | 14.0 % | 25.2 % | 37.6 % |
| BASRI-total | 14.7 % | 26.8% | 42.1 % |
| SASSS | 31.0 % | 38.1 % | 45.5 % |
| Modified SASSS | 41.6 % | 46.4% | 56.5 % |

BASRI = Bath Ankylosing Spondylitis Radiology Index; SASSS = Stoke Ankylosing Spondylitis Spinal Score.

# Discussion

In view of the results of our study we conclude that the modified SASSS seems the most appropriate method for scoring radiological progression in AS patients. This conclusion is based on the following aspects of the OMERACT filter:

*Truth*

A valid scoring system requires the assessment of the cervical and lumbar spine. Inclusion of the SI joints and hips has no additional value for the detection of progression. An AP view of the lumbar spine as well as an assessment of the posterior site of the lumbar spine do not provide sufficient additional information on progression to justify the extra effort, but an AP view will provide additional information (and therefore better reflects the truth) if the level of damage rather than the progression of damage is

the major concern. The consequences are that the SASSS is not recommended because it does not take into account the cervical spine. The BASRI is recommended because of its AP view if radiographic damage is the matter of interest, but this AP view does not supply valuable additional information if progression should be scored.

*Discrimination*

Concerning reliability the modified SASSS is superior with respect to interobserver reliability. With respect to sensitivity to change this method quantifies a higher proportion of patients as progressive as compared to the BASRI. It also appeared that BASRI in contrast with modified SASSS may suffer from a ceiling-effect.

*Feasibility*

The BASRI is taking less time for scoring and training but yields the highest radiation exposure for the patient. For the aspect of feasibility a method of preference is not revealed.

If the results of the comparison of the different scoring methods against the OMERACT filter are surveyed, the modified SASSS seems to be preferable for the evaluation of radiological progression in clinical trials and cohort studies

In the literature several studies are published about scoring of radiographs of AS patients, mainly by the developers of the BASRI and SASSS. It appears that these studies largely confirm our results.

In our study we could not find support for including the hips in a staging or a change score. This finding is supported by data of the group from Bath. Mackay et al.[1] explain persuasively why the hips are not included in the BASRI-spine. Because hip disease affects only 18-37% of the AS population the use of a global score for every AS patient, with a maximum score of 16 rather than 12, may inappropriately dilute the score of the majority of AS patients. Those with severe, or grade 4, spinal disease without hip arthritis would rate only 12 on a 16-point global scale despite having a bamboo spine, poor metrology and poor function. It may be better to grade these populations separately, using the BASRI-spine for one and the BASRI-total for the other. Note that omission of the hips and SI joints in our scoring method does not necessarily mean that these joints are not important in AS for prognostication. As an example, hip involvement turned out to be an important predictor of severe disease[14].

MacKay et al also support our conclusion about the essential inclusion of the cervical spine. They presented data on the involvement of cervical and lumbar spine, SI joints and hips in a group of 470 patients[15]. Over 80% of the patients showed involvement of the cervical and / or lumbar spine, or both (43%), 8% of the patients showed changes only in the cervical and not in the lumbar spine. Concerning the discussion which view is needed for scoring the lumbar spine, our conclusion on the status scores is again supported by Mackay et al.[1]. They judged 58 sets of AP and lateral views of the lumbar spine. These sets were scored on only the AP view, the lateral view and the

combination score (the highest score of the two views.) The combination score differed from the AP or lateral scores if syndesmophytes or fusion were seen at different levels on each projection. This occurred in 3 of the 58 patients. The 'combination' score differed from the AP score alone in 9 of the 58 patients (15.5%) and from the lateral score alone in 21 patients (36%). Overall, the use of 2 projections changed the score in 46% of the cases. Assuming that the combination view gives the most truthful assessment, the sensitivity for the AP view alone is 0.83 and that for the lateral view alone is 0.73. For the aim of staging, both views are therefore necessary. Unfortunately Mackay et al. did not investigate if both views are also necessary to assess progression.
In none of the scoring methods the thoracic spine is included. This is due to technical problems related to the anatomy of the chest with superimposed lung tissue. Another structure of the spine that has not been mentioned is the facet joints. In lateral views of the lumbar spine these joints are difficult to assess with any degree of confidence even by an experienced musculoskeletal radiologist[2]. On an AP view these joints can be assessed. This is an advantage of the BASRI scoring method. All other methods ignore the posterior structures of the spine, classifying those who have only posterior element fusion as normal or as having "mild" radiographic changes when in fact the spine may be completely fused[1]. In table 3.4 we compared measurements of spinal mobility with radiological scores. We found a good correlation, and this relation is as good for BASRI-spine as for the other methods.
An important disadvantage of the BASRI in comparison with the SASSS methods is the fact that it does not pick up minor radiological change. The score does not change with each additional erosion or sclerosis, and will always remain grade 2 or mild disease until there is fusion between 2 vertebrae or ≥3 syndesmophytes are identified. The developers of the BASRI and SASSS evaluated reliability and sensitivity to change. Inter- and intraobserver reliability of the BASRI was assessed on status scores[1], which showed a good reliability. After a period of one year no change was observed. In a two-year period the mean of BASRI spine increased statistically significant from 7.0 to 7.9 (in 40 patients). The radiographs in this study were blinded for the chronology, confirming that the BASRI could determine "forward progression" (i.e. could identify the earlier of 2 radiographs performed on the same individual). We found a progression from 6.5 to 6.9 over a 2-year time interval, and our films were read in chronological order, which often even amplifies progression scores. The difference might be explained by the fact that the patient population in Bath differs from the population in the OASIS cohort with respect to severity.
The developers of the SASSS[16] also investigated reliability of their method. They showed a good interobserver reliability but, unexpectedly, poorer intraobserver reliability. Sensitivity to change was assessed in 28 patients over a 12 months time interval, and the films were read in known order. The SASSS increased by 4.1 (from 14.4 to 18.4), which was statistically significant. This increase is considerable, in comparison with our results; after four years of observation we observed a progression of only 3.5 points.

The cohort used in this present study has been studied before, as mentioned in the introduction[5,6]. In contrast with the former study, in the present study a change after one year was observed but the order in which the radiographs were scored was known, while in the study of Spoorenberg et al. the order was unknown. This can markedly influence the results, as has been shown for RA[17]. Moreover, in the study of Spoorenberg the average of the two observers' progression scores were used to determine whether a patient was classified as being progressive, and the criteria to define progression were much stricter than those applied in the present study. This was especially a disadvantage for the (modified) SASSS. With the four-year data available we were able to observe that the minor changes after 1 and 2 years indeed forecasted further progression after 4 years, which adds to the validity of these minor changes.

The different results on progression of all studies can also be explained by difference in compilation of the patient populations. There are two different concepts on the mode of radiographic and functional progression of AS during the first 10 years after disease onset. While two groups[18,19] reported that the most rapid progression occurred in this period another group recently[20] reported that in their patient population radiographic progression was linear with no significant changes between the decades.

This study may evoke some concerns. First, the conclusions of this study are based on the OASIS cohort. Although this cohort represents the entire spectrum of AS patients, which adds to the external validity of the observations, the conclusions still need confirmation by other independent investigators in a different cohort. Second, it was not investigated whether any of the measures suffers from spectrum bias, e.g. whether it performs differently in patients with early- as compared to late AS. The group was too small to make subgroups for such an analysis. Third, most of the analyses in this study are based on the scores of one reader. Although interobserver reliability appeared to be satisfactory, future studies should include more readers in order to limit biases due to single observers.

In all studies describing measuring radiological change in AS patients there are no data available about reliability of progression scores, which is important in clinical trials. Therefore we would like to emphasize that in future studies it is necessary to pay attention to reliability of progression scores. As can be seen in our results, the reliability of progression scores can add important information to the reliability of status scores. In our study change could be assessed reliably by the modified SASSS.

In summary, comparing the BASRI, SASSS and modified SASSS with respect to their use in clinical trials, we have shown here that the modified SASSS offers advantages in measurement properties in comparison with the BASRI and SASSS.

However, the BASRI is a feasible and user-friendly method that reliably detects damage in patients with AS, and may be used for that purpose in clinical practice

# References


1. MacKay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI): a new, validated approach to disease assessment. Arthritis Rheum 1998;41: 2263-70.
2. Averns HL, Oxtoby J, Taylor HG, Jones PW, Dziedzic K, Dawes PT. Radiological outcome in ankylosing spondylitis: use of the Stoke Ankylosing Spondylitis Spine Score (SASSS). Br J Rheumatol 1996;35:373-6.
3. Creemers MC, Franssen MJ, van 't Hof MA, Gribnau FW, van de Putte LB, van Riel PL. A radiographic scoring system and identification of variables measuring structural damage in ankylosing spondylitis [thesis] Nijmegen (the Netherlands): University of Nijmegen; 1993.
4. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. Assessments in Ankylosing Spondylitis Working Group. J Rheumatol 1997;24:2225-9.
5. Spoorenberg A, de Vlam K, van der Heijde D, de Klerk E, Dougados M, Mielants H, van der Tempel H, Boers M, van der Linden S. Radiological scoring methods in ankylosing spondylitis: reliability and sensitivity to change over one year. J Rheumatol 1999;26: 997-1002.
6. Spoorenberg A, de Vlam K, van der Linden S, Dougados M, Mielants H, van de Tempel H, van der Heijde D. Radiological scoring methods in Ankylosing Spondylitis. Reliability and sensitivity to change over one and two years. J Rheumatol 2004;31:125-32.
7. Boers M, Brooks P, Strand CV, Tugwell P. The OMERACT filter for Outcome Measures in Rheumatology. J Rheumatol 1998;25:198-9.
8. Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. Ankylosing spondylitis: the correlation between a new metrology score and radiology. Br J Rheumatol 1995;34:767-70.
9. Dale K. Radiographic gradings of sacroiliitis in Bechterew's syndrome and allied disorders. Scand J Rheumatol 1979;32(S 32):92-7.
10. MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in ankylosing spondylitis: the bath ankylosing spondylitis radiology hip index. J Rheumatol 2000;27:2866-72.
11. Taylor HG, Beswick EJ, Dawes PT. Sulphasalazine in ankylosing spondylitis. A radiological, clinical and laboratory assessment. Clin Rheumatol 1991;10:43-8.
12. Bellamy N: Musculoskeletal clinical metrology. Dordrecht; Kluwer Academic Publishers Group 1993:25-9.
13. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, Jenkinson T. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281-5.
14. Amor B, Santos RS, Nahal R, Listrat V, Dougados M. Predictive factors for the longterm outcome of spondylarthropathies. J Rheumatol 1994;21:1883-7
15. Mackay K Brophy S, Mack C, Calin A. Patterns of radiological axial involvement in 470 ankylosing spondylitis patients [abstract]. Arthritis Rheum 1997;40 (S 9):S61.
16. Taylor HG, Wardle T, Beswick EJ, Dawes PT. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. Br J Rheumatol 1991;30: 330-5.
17. Bruynesteyn K, van der Heijde D, Boers M, SAudan A, Peloso P, Paulus H, Houben H, Griffiths B, Edmonds J, Bresnihan B, Boonen A, Van Der Linden S. Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. J Rheumatol 2002;29:2306-12.

18. Carette S, Graham D, Little H, Rubenstein J, Rosen P. The natural disease course of ankylosing spondylitis. Arthritis Rheum 1983;26:186-90.
19. Gran JT, Skomsvoll JF. The outcome of ankylosing spondylitis: a study of 100 patients. Br J Rheumatol 1997;36:766-71.
20. Brophy S, Mackay K, Al-Saidi A, Taylor G, Calin A. The natural history of ankylosing spondylitis as defined by radiological progression. J Rheumatol 2002;29:1236-43.

# Chapter 4

# Scoring of radiographic progression in randomized clinical trials in ankylosing spondylitis: a preference for paired reading order

A Wanders, R Landewé, A Spoorenberg, K de Vlam, H Mielants, M Dougados, Sj van der Linden, D van der Heijde

*Ann Rheum Dis 2004 Published Online First [Aug 5, 2004]*

# Abstract


Objective

To describe the influence of the reading order (chronological vs. paired) on radiographic scoring results in Ankylosing Spondylitis (AS). Since paired reading is requested for establishing drug efficacy in clinical trials, we investigated whether this method is sufficiently sensitive to change.

Methods

Films obtained from 166 patients (at baseline, 1 year and 2 years) were scored by one observer, using the modified Stoke Ankylosing Spondylitis Spinal Score. Films were first scored chronological, and were scored paired 6 months later.

Results

Reading chronological yields more radiographic progression; both at 1 year (mean progression: 1.3 (standard deviation (SD) 2.6) vs. 0.5 (SD 2.4) units) and at 2 years (mean progression: 2.1 (SD 3.9) vs. 1.0 (SD 2.9) units). Reading with chronological order yielded significantly more progression than paired reading (between- method difference: $p<0.001$ at 1 year, and $p<0.001$ at 2 years). After 1 year, progression (>0 units) was found in 21% of patients after paired reading and in 33% after chronological reading. After 2 years, these numbers were 30% and 41%, respectively. Sample size calculations showed that 94 patients per treatment arm are required in a randomized clinical trial (RCT) to provide sufficient statistical power to detect a difference in 2-year progression if films are scored paired.

Conclusion

Reading with chronological time order is more sensitive to change than reading with paired time order, but paired reading is sensitive enough to pick up change with a follow-up of 2 years, resulting in an acceptable sample size for RCTs.

# Introduction

For evaluation of therapy in Ankylosing Spondylitis (AS), the ASsessment in Ankylosing Spondylitis (ASAS) working group has developed core sets to be used in various settings(1), including the setting for disease controlling antirheumatic therapy (DC-ART). One segment of the definition of DC-ART reads: "prevent or significantly decrease the rate of progression of structural damage"[1]. To assess progression of structural damage, radiographic outcome assessment is included in the DC-ART core set. Radiology as outcome parameter in AS clinical trials is new, in contrast to clinical trials in rheumatoid arthritis (RA) in which radiographic outcome already has a prominent place. The methodology of radiographic scoring in AS is still developing. Recently, we performed a study comparing the existing radiographic scoring methods with respect to various aspects of validity[2]. It was concluded that the modified Stoke Ankylosing Spondylitis Spinal Score (modified SASSS) is the most appropriate method for use in clinical trials.
It is known from studies concerning evaluation of radiographic damage in RA that the order in which films are presented to the observer influences results[3-6]. Films can be grouped per patient and presented to the reader without knowing the chronological order of the films: paired scoring. Films can also be grouped per patient and presented in chronological order. The advantage of chronological reading is that it provides the reader with a maximum of information, thereby reducing 'true' measurement error. Reading films chronologically results in an increased ability to detect changes as compared to paired reading. In 1999 van der Heijde et al. showed that reading with chronological order was more sensitive to change than paired reading in RA[5]. However, the possibility that chronological reading overestimated progression of joint damage because readers expected to see progression (expectation bias) could not be excluded. In a follow-up study Bruynesteyn et al.[6] used progression considered as clinically relevant by rheumatologists as a proxy for true progression, and concluded that paired reading underestimates the true progression. The advantage of paired reading, however, is that expectation bias is almost ruled out: readers are not aware of the sequence of the films and therefore do not tend to score more progression in the follow-up film. The issue which of both reading orders should be used is therefore not unanimously answered by the above-mentioned studies. Despite this controversy, the reading of structural damage in RA clinical trials is predominantly performed by readers blinded for the sequence. This stems from the general epidemiological consensus that in order to prevent bias observers must be blinded as far as possible, and from the practical aspect that for registration purposes reading with blinded sequence is requested by the drug regulatory agencies. Therefore it seems obvious that radiographic progression in AS clinical trials should also be assessed by paired reading. However, there is some concern that paired scoring in AS is not sensitive enough, since progression occurs slowly, and only in a minority of patients[7]. The aim of this study therefore was 1) to explore the differences with respect to sensitivity to change between paired and chronological scoring in AS, and 2) to investigate whether trials with

radiographic progression as primary endpoint can be designed, that have sufficient statistical power with feasible patient numbers if films are read with paired order.

# Methods

### Patients and films

Radiographs from an international longitudinal, observational study on outcome in AS, the OASIS cohort, were used[8]. Originally 217 patients from four centers in the Netherlands, Belgium and France were included in this cohort. Radiographs were obtained at baseline, and after one and two years of follow-up. After two years of follow-up, complete sets of radiographs of baseline, one and two year of 166 patients were available; only these patients were included in this study. The modified SASSS was assessed on lateral views of the lumbar and cervical spine.

### The scoring of films.

The modified SASSS method scores every corner of the anterior site of the lumbar and cervical vertebrae on a scale from 0–3, in which 0 indicates no abnormalities, 1 is used for erosion, sclerosis or squaring, 2 indicates a syndesmophyte and 3 a bridging syndesmophyte. This yields a possible total score of 72 units. The lumbar spine is scored from the lower border of the $12^{th}$ thoracic vertebra to the upper border of the first sacral vertebra, the cervical spine is scored from the lower border of the second cervical vertebra to the upper border of the first thoracic vertebra. In a previous study was shown that this method had a good inter- and intraobserver reliability[2]. Intraclass correlation coefficients for inter- and intraobserver reliability for progression scores with a 2-year interval were 0.82 and 0.95, respectively. Films were available for three time points: baseline, 1 year and 2 years. First the films were scored in chronological order, and after 6 months the films were scored again by the same reader (AW), but now in a random time order (paired films per patient). The chronological scoring method allows negative progression scores.

### Analysis and statistics

Descriptive statistics (mean, standard deviation, median, $25^{th}$ and $75^{th}$ percentile) are given for the modified SASSS scores for both reading orders at the three time points, as well as for the progression scores. Also are descriptive statistics provided for those patients who had a radiographic progression greater than zero. To visualize the effects of scoring by the two reading orders, progression scores obtained by both methods were plotted by its cumulative frequency (expressed as percentage; cumulative probability) in probability plots[9]. Wilcoxon's signed ranks test was used to test the null-hypothesis that 1- or 2-year progression is zero. Mann-Whitney test was used to investigate the null hypothesis that radiographic progression obtained by both reading

orders was similar. Proportions of patients with progression (>0 units) by reading order at 1- or 2 years were compared by chi-square test. Sample sizes for a putative RCT with one untreated control group and one active treatment group, and radiographic progression as primary endpoint, were calculated using the power calculator of the University of California, Los Angeles (http://calculators.stat.ucla.edu/powercalc/ , significance level = 0.05, 2-sided, power = 0.80). This was done under the assumptions that an untreated control group will show progression as in the OASIS cohort, and that progression in the active treatment is zero, with a standard deviation equal to the standard deviation in the untreated control group. Van der Waerden-normalized progression scores were used to perform the sample size calculations.

# Results

## Sensitivity to change

Table 4.1 shows the patient characteristics at baseline.

Table 4.1 Patient characteristics at baseline

| | mean | SD | median | p25 | p75 |
|---|---|---|---|---|---|
| Age (years) | 43.9 | 12.5 | 43.1 | 33.6 | 52.9 |
| Mean duration of complaints (years) | 20.4 | 11.6 | 17.1 | 12.0 | 27.5 |
| Mean duration of disease after diagnosis (years) | 11.7 | 9.0 | 10.0 | 4.8 | 15.4 |
| Male, % | 71.5 | | | | |

In table 4.2 the descriptive statistics of the modified SASSS scores according to chronological and paired reading are given.
Baseline scores are almost similar. Reading with chronological order yields more progression than paired reading, both at 1 year 1.3 (2.6) (mean (standard deviation)) units vs. 0.5 (2.4) units, and at 2 years 2.1 (3.9) vs. 1.0 (2.9) units.
Table 4.3 provides the descriptive statistics of the modified SASSS scores of those patients who showed a progression greater than zero.

Table 4.2 Descriptive statistics of 1-year and 2-year follow-up of radiological damage scored according to the modified SASSS with paired and chronological reading order ( n = 166 patients).

| | mean | SD | median | p25 | p75 |
|---|---|---|---|---|---|
| *Paired reading order* | | | | | |
| Baseline | 13.1 | 18.0 | 4.9 | 0.0 | 18.1 |
| 1 year | 13.6 | 18.4 | 4.9 | 0.0 | 21.2 |
| 2 years | 14.1 | 18.6 | 6.0 | 0.0 | 23.3 |
| Progression after 1 year | 0.5 | 2.4 | 0.0 | 0.0 | 0.0 |
| Progression after 2 years | 1.0 | 2.9 | 0.0 | 0.0 | 1.6 |
| *Chronological reading order* | | | | | |
| Baseline | 13.6 | 19.2 | 5.0 | 0.0 | 17.0 |
| 1 year | 14.9 | 19.7 | 6.0 | 0.0 | 20.1 |
| 2 years | 15.8 | 20.1 | 6.0 | 0.0 | 22.7 |
| Progression after 1 year | 1.3 | 2.6 | 0.0 | 0.0 | 1.4 |
| Progression after 2 years | 2.1 | 3.9 | 0.0 | 0.0 | 3.0 |

Modified SASSS = modified Stoke Ankylosing Spondylitis Spinal Score; SD = standard deviation; p25 = 25$^{th}$ percentile; p75 = 75$^{th}$ percentile.

Table 4.3 Descriptive statistics of 1-year and 2-year follow-up of radiological damage scored according to the modified SASSS of the patients who showed progression greater than zero accordingly to the paired and chronological reading order.

| | n | mean | SD | median | p25 | p75 |
|---|---|---|---|---|---|---|
| *Paired reading order* | | | | | | |
| Progression after 1 year | 35 | 4.0 | 2.9 | 4.0 | 1.6 | 5.0 |
| Progression after 2 years | 50 | 3.9 | 3.8 | 2.9 | 2.0 | 5.0 |
| *Chronological reading order* | | | | | | |
| Progression after 1 year | 55 | 3.9 | 3.1 | 3.0 | 1.2 | 6.0 |
| Progression after 2 years | 68 | 5.2 | 4.6 | 4.0 | 2.0 | 7.0 |

In the entire cohort of this study, both methods picked up progression from baseline significantly (chronological order: $p<0.001$ for 1 year, and $p<0.001$ for 2 years; paired order: $p=0.021$ for 1 year, and $p<0.001$ for 2 years). Reading with chronological order was significantly more sensitive than paired reading (between-method difference: $p<0.001$ at 1 year, and $p<0.001$ at 2 years). After 1 year of follow-up 21% of patients showed progression >0 units according to the paired reading results and 33% of patients according to the chronological reading results. At 2-year follow-up these numbers were 30% and 41%, respectively.

This progression pattern is further illustrated by probability plots for the 1-year interval (figure 4.1) and for the 2-year interval (figure 4.2). In figure 4.1 it can be seen that for both scoring method a majority of patients do not show progression. This was already represented by the median that was zero for both methods (this median value can be found on the x-axis at a proportion percentage of 0.50) The advantage of the probability

plot is that it also easily represents the percentage of patients with progression: For instance, figure 4.1 for the chronological reading order shows that the curve deviates from zero at a value of 67%, indicating that 33% of patients show progression. Although negative progression scores were allowed in the chronological scoring method, these are not seen in the two plots. For the paired scoring method negative scores are visible in both figures (15% at 1 year and 11% at 2 years).
A comparison of both plots shows that the curve for chronological reading lies most left, which indicates that with the chronological reading more patients are qualified as progressive. The difference between these two curves was statistically tested; both for the 1-year interval and for the 2-year interval the difference between both methods was statistically significant (p=0.019 and p=0.051 respectively).

Sample size calculations for the paired reading order

In table 4.1 and in the probability plots it is shown that the data of the paired scoring order have a skewed distribution. So before entering the data in sample size calculations a van der Waerden normalization procedure was performed. The following assumptions were made in the sample size calculations: the mean progression in the intervention group is zero and the standard deviation is the same as in the control group (the OASIS cohort). With these assumptions the following sample sizes were obtained for a RCT in which radiographic progression is scored according to the modified SASSS by paired reading order; A RCT with a duration of 1 year requires 922 patients per arm, and a RCT with a duration of 2 years requires 94 patients per arm, in order to statistically underscore a true between-group difference of 0.5 units (1 year) respectively 1.0 units (2 years).

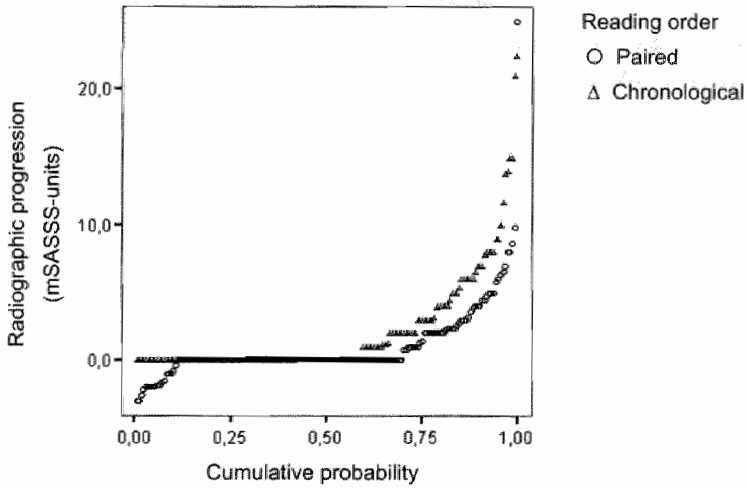



Figure 4.1 Probability plot of 1-year progression in modified SASSS scores for paired and chronological reading order.

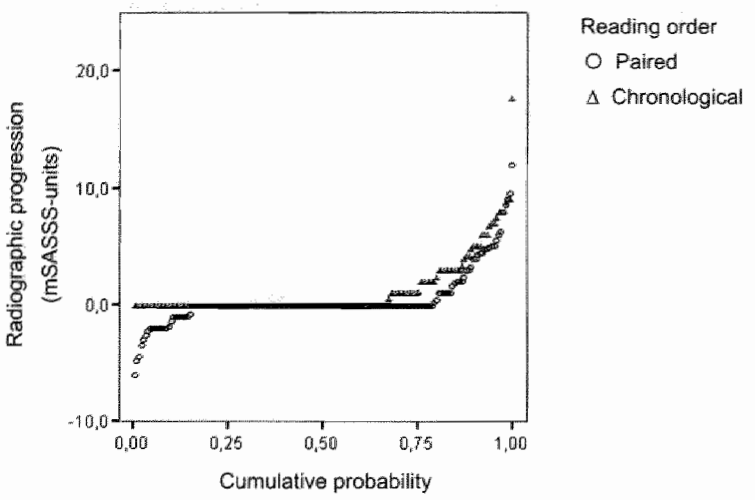



Figure 4.2 Probability plot of 2-year progression in modified SASSS scores for paired and chronological reading order.

# Discussion

The conclusion of this study is that the order by which films of AS patients are presented to the observer influences the reading results, which is in accordance with the findings in RA. Reading films in chronological order shows a higher mean progression and a greater proportion of patients with progression, in comparison with paired reading.
However, we also showed that scoring with a paired reading order is sufficiently sensitive to pick up radiographic progression after 2 years of follow up, under the specific conditions set in this study. To illustrate the feasibility of paired scoring in trials with a radiographic endpoint, we demonstrated an acceptable sample size for a putative RCT, using real progression data from the OASIS cohort, provided that the duration of the trial is at least 2 years.

The theoretical assumption that chronological scoring in comparison to paired scoring would have a higher sensitivity to change, which is supported by data from research in RA, was confirmed in this study. It was also seen that the magnitude of the signal picked up by the chronological reading order is greater than by the paired reading order. However, which part of this signal is a 'true' effect and which part can be attributed to 'noise' is difficult to establish, especially for the chronological reading order. First of all, in the chronological reading order expectation bias contributes to 'noise', whereas this

bias is almost ruled out in paired scoring. It is impossible to determine in chronological reading which part of 'noise' is caused by expectation bias and which part by the remaining measurement error. The measurement error in paired reading can be visualized by means of probability plots. Since it is thought that the phenomenon of healing ("true negative scores") does not occur in AS (which is supported by the results of chronological reading data, in which no negative scores were found), the negative scores by paired scoring can be considered as measurement error. When this is applied to figure 4.1 with a 1-year time interval, than it is seen that 15 % of the patients have a negative score. The percentage of patients that have a positive score is 21%. Assuming that measurement error works equally in both directions, this would mean that only 6% of patients show 'real' progression. In figure 4.2, with a 2-year interval, it is seen that 11% of patients have a negative score and 30% of patients have a positive score, which means that 19% of patients show 'real' progression. This difference in signal-noise ratio is also reflected by the sample size calculations, after one year of follow-up a huge sample size is needed, 922 patients, versus 94 patients for a follow-up of 2 years.

The lack of expectation bias and the possibility of assessing measurement error are advantages of paired reading. Apart from these advantages, it is also a fact that this scoring method is requested by the agencies for registration purposes. Therefore the feasibility of this method is relevant with respect to the number of patients needed to demonstrate a significant difference in radiographic progression. The problem with sample size calculations is that they are dependent on the assumptions, which are arbitrary. Determining the assumptions underlying a RCT with radiographic progression in AS as outcome parameter is particularly difficult, because not much is known about the effect of interventions on radiographic progression. Data from a study in RA[10] showed that anti-TNF treatment inhibited radiographic progression, which might be supportive for our assumption of a progression of zero. However, despite all the assumptions and uncertainties associated with sample size calculations, there is a precedent that shows that a sample size of 94 patients may provide sufficient statistical power. Recently a RCT in AS was performed in which radiographic progression of 2-years was used as primary outcome parameter[11]. In this RCT continuous versus on demand intake of non-steroidal-anti-inflammatory drugs was compared with regard to radiographic progression. Radiographic progression was assessed by the modified SASSS with a paired scoring order. The two treatment groups consisted of 74 and 76 patients, and a between-group difference of 1.1 was found to be statistically significant in this study.

Therefore, based on theoretical arguments and on the results of this study we recommend RCTs in AS with radiographic progression as an endpoint to be designed with 2 years duration and to be scored by paired reading order.

# References


1. van der Heijde D, van der Linden S, Bellamy N, Calin A, Dougados M, Khan MA. Which domains should be included in a core set for endpoints in ankylosing spondylitis? Introduction to the ankylosing spondylitis module of OMERACT IV. J Rheumatol 1999;26:945-7.
2. Wanders AJB, Landewé RBM, Spoorenberg A, Dougados M, van der Linden S, Mielants H, van der Tempel H, van der Heijde DM. What is the most appropriate radiologic scoring method in Ankylosing spondylitis clinical trials. A comparison based on the OMERACT filter. Arthritis Rheum 2004; 50:2622-32.
3. Ferrara R, Priolo F, Cammisa M, Bacarini L, Cerase A, Pasero G, Ferraccioli GF, Alberighi OD, Antonellini A, Marubini E. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the GRISAR Study. Gruppo Reumatologi Italiani Studio Artrite Reumatoide. Ann Rheum Dis 1997;56:608-12.
4. Salaffi F, Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. J Rheumatol 1997;24:2055-6.
5. van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology (Oxford) 1999;38:1213-20.
6. Bruynesteyn K, van der Heijde D, Boers M, Saudan A, Peloso P, Paulus H, Houben H, Griffiths B, Edmonds J, Bresnihan B, Boonen A, Van Der Linden S. Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. J Rheumatol 2002;29:2306-12.
7. Spoorenberg A, de Vlam K, van der Linden S, Dougados M, Mielants H, van de Tempel H, van der Heijde D. Radiological scoring methods in Ankylosing Spondylitis. Reliability and sensitivity to change over one and two years. J Rheumatol 2004; 31:125-32.
8. Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. J Rheumatol 1999;26: 980-4.
9. Landewé R, van der Heijde DFMF. Radiographic progression depicted by probability plots: presenting data with optimal use of individual values. Arthritis Rheum 2004;50:699-706.
10. Lipsky PE, van der Heijde DM, St Clair EW, Furst DE, Breedveld FC, Kalden JR, Smolen JS, Weisman M, Emery P, Feldmann M, Harriman GR, Maini RN; Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. N Engl J Med 2000;343: 1594-602.
11. Wanders A, van der Heijde D, Landewé R, Béhier J-M, Calin A, Olivieri I, et al. Inhibition of radiographic progression in ankylosing spondylitis by continuous use of NSAIDs. ACR 2003;abstract 518.

# Chapter 5

# The association between radiographic damage of the spine and spinal mobility for individual patients with ankylosing spondylitis: Can the assessment of spinal mobility be a proxy for radiographic evaluation?

A Wanders, R Landewé , M Dougados, H Mielants, Sj van der Linden, D van der Heijde

*Ann Rheum Dis 2005: accepted for publication*

# Abstract

Objective
To demonstrate the association between various measures of spinal mobility and radiographic damage of the spine in individual patients with Ankylosing Spondylitis, and to determine whether the assessment of spinal mobility can be a proxy for the assessment of radiographic damage.

Methods
Radiographic damage was assessed by the mSASSS. Cumulative probability plots combined individual's radiographic damage score with the corresponding score for nine spinal mobility measures. ROC-analysis was performed to determine the cut-off level of every spinal mobility measure that discriminates best between the presence and absence of radiographic damage. Three arbitrary cut-off levels for radiographic damage were investigated. Likelihood ratios were calculated in order to further explore the diagnostic properties of the spinal mobility measures.

Results
Cumulative probability plots demonstrated the association between spinal mobility measures and radiographic damage for the individual patient. Irrespective of the chosen cut-off level for radiographic progression, lateral spinal flexion and Bath Ankylosing Spondylitis Metrology Index performed best with respect to discrimination between patients with- and those without structural damage. However, even the best discriminatory spinal mobility assessments misclassified a considerable proportion of patients (up to 20%). Intermalleolar distance performed worst (up to 30% misclassifications). Lateral spinal flexion performed best in predicting the absence of radiographic damage and modified Schober performed best in predicting the presence of radiographic damage.

Conclusion
This study unequivocally demonstrates the relationship between spinal mobility and radiographic damage. However spinal mobility can not be used as a proxy for radiographic damage in an individual patient.

# Introduction

The hypothesis that radiographic damage of the spine in patients with Ankylosing Spondylitis (AS) is associated with impairment of spinal mobility is confirmed by several studies[1-4]. However these studies investigated the relationship on a group level. The association between structural damage and various instruments to assess spinal mobility in the individual patient has never been reported to our knowledge. Given the fact that radiographic evaluation is a burden for the patient (radiation exposure), physician (time-consuming) and society (cost aspect), it might be relevant to investigate whether assessment of spinal mobility can be used as a proxy for the assessment of radiographic damage in individual patients.
The aim of this study therefore is twofold: 1) to demonstrate the association between various measures of spinal mobility and radiographic damage of the spine in individual patients and 2) to determine whether the assessment of spinal mobility can be used to replace radiographic evaluation of the spine.

# Methods

## Patients

This study was performed in the OASIS cohort, an international, observational study on outcome in AS which has been described in detail before[5]. Originally 217 consecutive outpatients from four centers in the Netherlands, Belgium and France were included in this cohort. Patient characteristics concerning demographic data, radiographic damage and spinal mobility measures are presented in table 5.1.

## Films

Films were scored by the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS)[6] by one observer (AW). In previous studies it was shown that this method had a good inter- and intraobserver reliability[6-8] and showed good discriminatory properties[8].

## Spinal mobility measures

Nine measures were performed: chest expansion, finger-to-floor distance, occiput to wall distance, tragus to wall distance, modified Schober, lateral spinal flexion, cervical rotation: intermalleolar distance, Bath Ankylosing Spondylitis Metrology Index (BASMI): this index is calculated using cervical rotation, tragus to wall distance, lateral spinal flexion, modified Schober and intermalleolar distance[9]. Each of the 5 BASMI measurements is divided into 11 equal sections[10], the mean of the 5 scores producing a BASMI score from 0.0 to 10.0. For all measures, the best of two attempts is recorded, and rounded at 0.1 cm, except for cervical rotation (1 degree) and BASMI (rounded at one decimal).

Table 5.1 Patient characteristics of the OASIS cohort at baseline (n = 199)

| Variable | mean | SD | median | P25 | P75 | min | max |
|---|---|---|---|---|---|---|---|
| Age (years) | 43.7 | 12.7 | 43.1 | 33.4 | 52.9 | 19.0 | 78.0 |
| Duration of complaints (years) | 20.1 | 11.7 | 17.1 | 11.7 | 27.1 | 0.0 | 52.5 |
| Duration of disease after diagnosis (years) | 11.4 | 9.0 | 9.5 | 4.3 | 15.2 | 0.2 | 42.0 |
| Male (%) | 71 | | | | | | |
| Radiographic changes | | | | | | | |
| mSASSS (units) | 14.0 | 19.5 | 5.0 | 0.0 | 18.0 | 0.0 | 72.0 |
| Spinal mobility measures | | | | | | | |
| Chest expansion (cm) | 4.7 | 2.2 | 4.5 | 3.2 | 6.0 | 0.4 | 12.5 |
| Finger to floor distance (cm) | 14.2 | 13.5 | 12.2 | 1.0 | 22.9 | 0.0 | 56.5 |
| Occiput-to-wall distance (cm) | 3.8 | 5.4 | 1.0 | 0.0 | 6.0 | 0.0 | 26.0 |
| Tragus-to-wall distance (cm) | 14.1 | 4.4 | 12.5 | 11.1 | 15.8 | 0.8 | 32.5 |
| Modified Schober (cm) | 2.9 | 4.4 | 3.0 | 1.8 | 4.0 | 0.0 | 6.8 |
| Lateral spinal flexion (cm) | 11.2 | 5.7 | 10.9 | 6.4 | 15.3 | 1.2 | 26.1 |
| Cervical rotation (degrees) | 64.2 | 23.1 | 68.0 | 51.0 | 81.0 | 6.0 | 107.0 |
| Intermalleolar distance (cm) | 105.0 | 21.7 | 105.1 | 94.0 | 119.0 | 38.0 | 152.0 |
| BASMI | 3.6 | 1.6 | 3.5 | 2.4 | 4.4 | 1.0 | 8.0 |

OASIS = Outcome in Ankylosing Spondylitis International Study; mSASSS = modified Stoke Ankylosing Spondylitis Spinal Score; BASMI = Bath Ankylosing Spondylitis Metrology Index; SD = standard deviation; P25 = 25$^{th}$ percentile; P75 = 75$^{th}$ percentile; min = minimum; max = maximum.

## Analysis

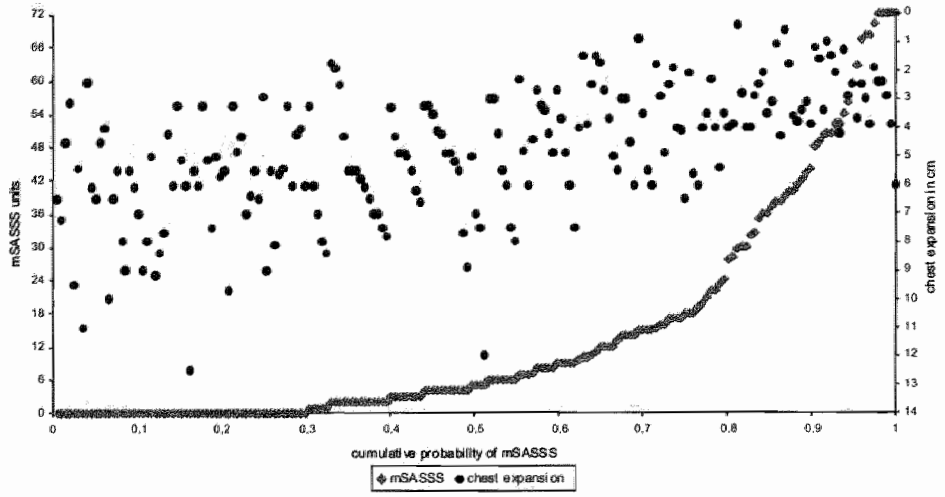
To demonstrate an association between spinal mobility and radiographic damage, combined scatter and cumulative probability plots were created for spinal mobility with radiographic damage. These plots combine every individual radiographic damage score with the corresponding score for each of the nine spinal mobility measures. The individual mSASSS scores of all patients is plotted by its cumulative order (from the lowest value starting at zero to the highest values ending at 100%). The combined procedure yields a scatterplot (observations of two variables combined), in which the values of one of the variables (mSASSS) is plotted against it cumulative frequency. Correlations on a group level were expressed as Spearman's rho. Receiver operating characteristic (ROC)-analysis was performed to determine the cut-off level for every spinal mobility measure that discriminates best (highest accuracy) between the presence and absence of radiographic damage. Three arbitrary cut-off levels for radiographic damage were investigated: 0, 3 and 6 mSASSS units. Sensitivity was considered the ability of every spinal mobility measure to truly indicate radiographic damage. Specificity was considered the ability of every spinal mobility measure to truly indicate the absence of radiographic damage. The area under the ROC-curve (AUC-ROC) was supposed to represent the discriminatory power of the spinal mobility

measure (an AUC-ROC of 0.5 means "no discriminatory power" and an AUC-ROC of 1.0 means "ideal discriminatory power"). Likelihood ratios (LR) for a positive test result (abnormal spinal mobility measure) (LR+) and for a negative test result (normal spinal mobility measure) (LR-) were calculated to further explore the diagnostic properties of individual and combined (BASMI) spinal mobility measures. To investigate whether radiographic damage could be predicted accurately on the basis of spinal mobility measures, post-test probabilities on the absence or presence of radiographic damage were calculated making use of Bayes theorem (post-test odds on radiographic damage = LR+ * pre-test odds on radiographic damage) [11].

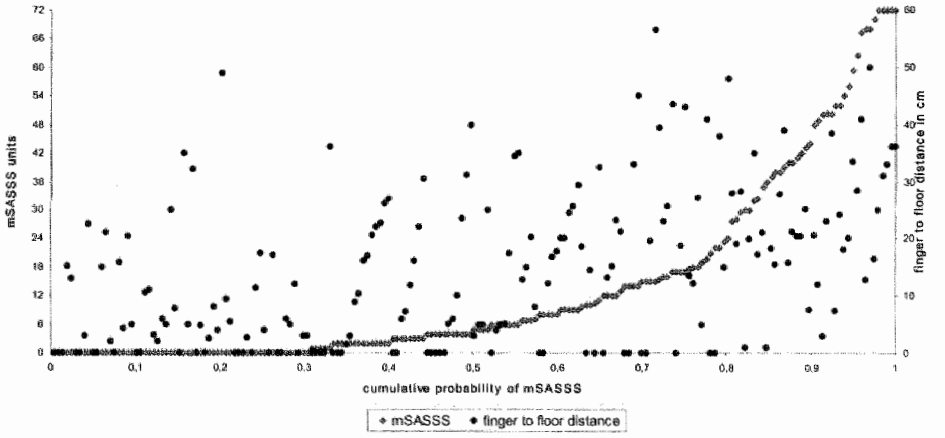
# Results

The patient characteristics, as well as the radiographic damage scores and the scores for all nine spinal mobility measures and the BASMI are presented in table 5.1. Of the total cohort of 217 patients, complete data sets were available for 199 patients. For the purpose of this study, it is important that the individual values of both the radiographic as well as the spinal mobility measures cover a range that is as broad as possible. It can be seen from table 5.1 that OASIS includes both patients with normal spinal mobility and absence of radiographic damage, as well as patients with severely impaired levels of spinal mobility and high degrees of radiographic structural damage. The distribution pattern of all spinal mobility measures can be deducted from interpreting the mean and median values.

The detailed scatterplots of the individual mSASSS scores in cumulative order versus the nine spinal mobility measures are presented in figure 5.1. In order to better illustrate the information that is provided by this type of scatterplot, we discuss the relation between lateral spinal flexion and mSASSS as an example (figure 5.1g).The probability plot of the mSASSS scores (squared symbols) visualizes that 30% of the patients has a mSASSS of 0 units, that the median mSASSS score is 5.0, and that a minority of patients have very high scores, reaching up to the maximum of 72 units in a few patients. Every mSASSS score is combined with the corresponding value for lateral spinal flexion (round symbols) in that patient (one x-axis value has two y-axis values) that can be read from the second y-axis. Eyeballing the distribution pattern of the lateral spinal flexion scores now learns that the pattern converges from wide (high level of dispersion) in case of normal mSASSS scores (left side of the graph) to narrow (low level of dispersion) in case of the highest mSASSS scores (right side of the graph). Or rephrased: lateral spinal flexion can range from entirely normal to highly abnormal if radiographic damage is absent, but lateral spinal flexion is almost always abnormal if radiographic damage is severe. Some of the spinal mobility measures (chest expansion, finger-to-floor distance) more or less reflects the same kind of relationship. Other spinal mobility measures, however, shows different relationships. Tragus-to-wall- and occiput-to-wall-distance (twice the same concept) do not converge: absence of radiographic damage does not rule out abnormal spinal mobility, and vice versa.

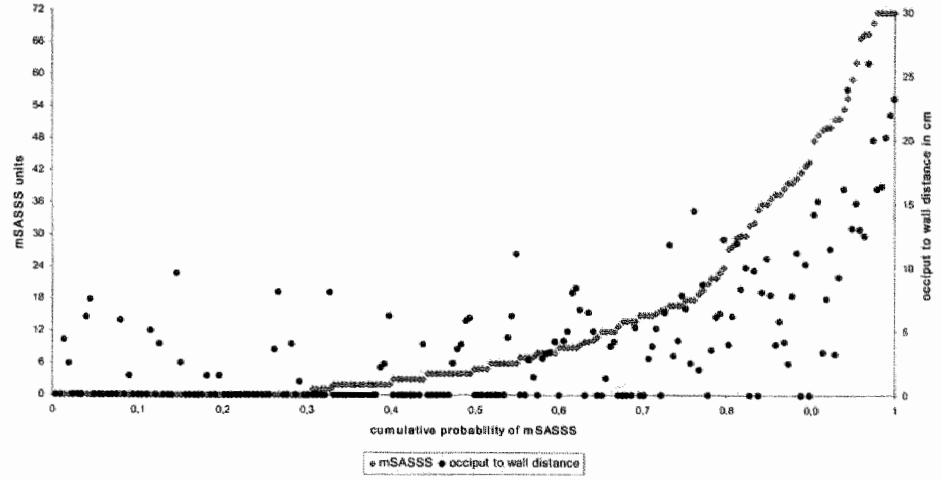



1 a cumulative mSASSS versus chest expansion

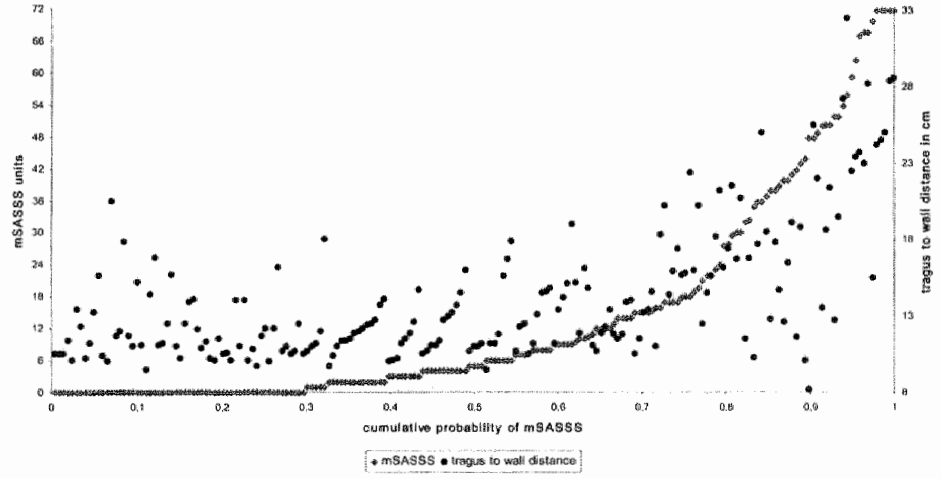



1 b cumulative mSASSS versus finger to floor distance

Figure 5.1 Scatter plots of cumulative mSASSS versus spinal mobility measures

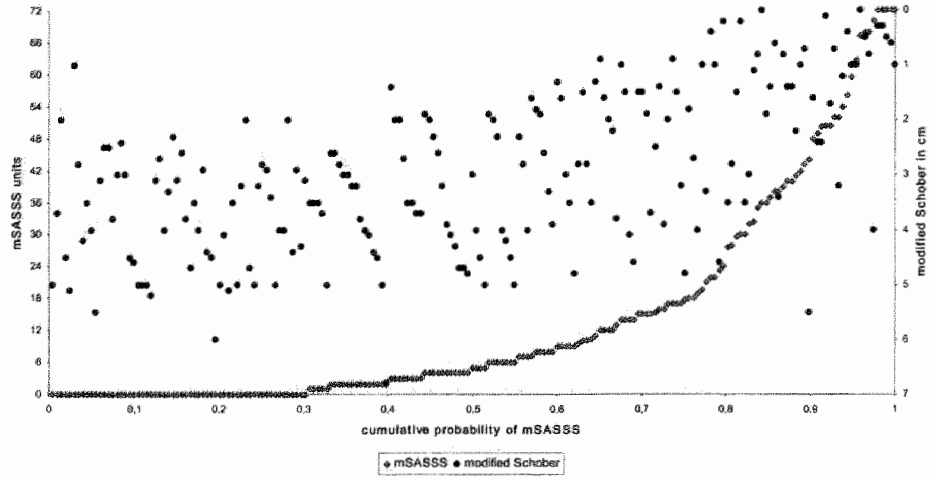



1 c cumulative mSASSS versus occiput to wall distance

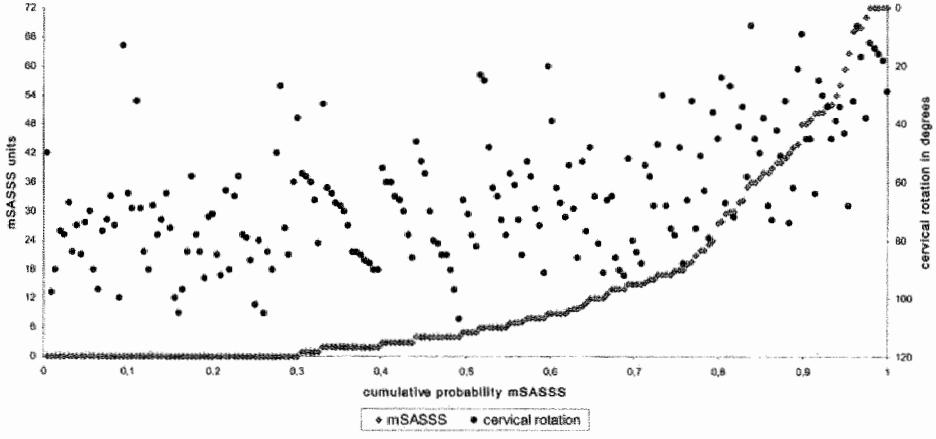



1 d cumulative mSASSS versus tragus to wall distance

Figure 5.1 Scatter plots of cumulative mSASSS versus spinal mobility measures

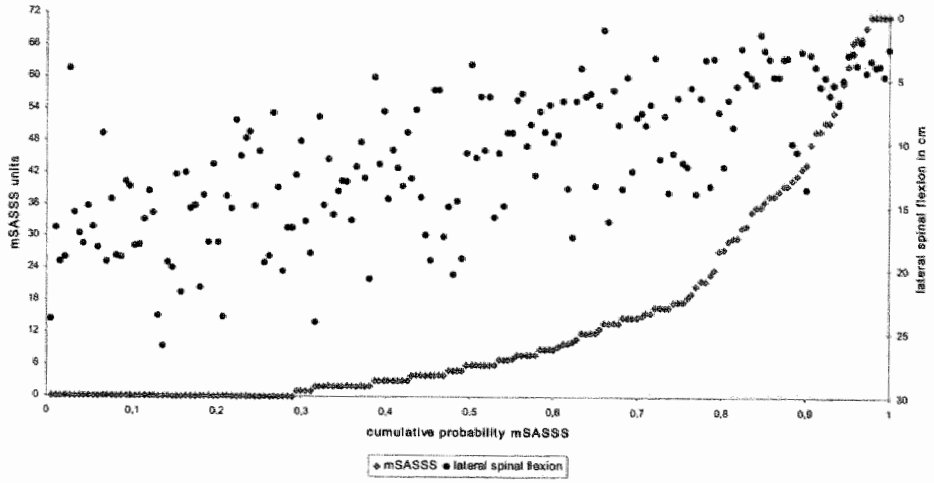



1 e cumulative mSASSS versus modified Schrober

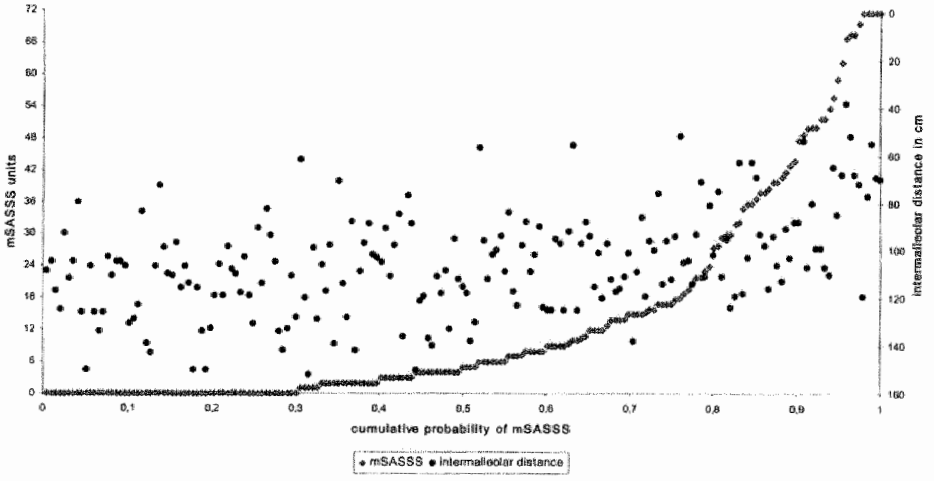



1 f cumulative mSASSS versus cervical rotation

Figure 5.1 Scatter plots of cumulative mSASSS versus spinal mobility measures

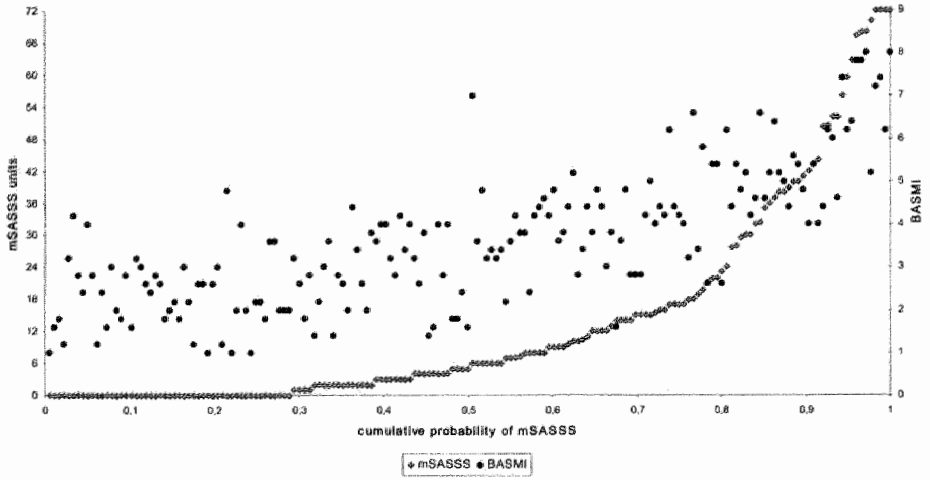



1 g cumulative mSASSS versus lateral spinal flexion




1 h cumulative mSASSS versus intermalleolar distance

Figure 5.1 Scatter plots of cumulative mSASSS versus spinal mobility measures




1 i cumulative mSASSS versus BASMI

Figure 5.1 Scatter plots of cumulative mSASSS versus spinal mobility measures

Only in case of a very high mSASSS score, both spinal mobility measures are obligatory impaired. Intermalleolar distance and Schober's test have a somewhat different pattern: highly abnormal levels are lacking in case of absence of damage, and in turn normal levels are lacking in case of a high level of damage.

Altogether, the data suggest that spinal mobility assessed with various instruments can be impaired by radiographic damage as well as by other unrelated processes, and the relation between impaired spinal mobility is only strong in case of a high level of radiographic damage.

The statistically significant Spearman's correlation coefficients on a group level presented in table 5.2 with correlations ranging from –0.42 (intermalleolar distance) to 0.76 (BASMI ), also informs about an association between spinal mobility and radiographic damage. In addition also the correlations of the separate components (cervical and lumbar) of the mSASSS with the spinal measures are presented in table 5.2. As could be expected the correlation of the part that corresponds with the spinal mobility measure is higher than the correlation of the part that does not correspond, for instance: the modified Schober measures the mobility of the lower lumbar spine, correlation with mSASSS lumbar r = -0.64 versus mSASSS cervical r = -0.40.

Table 5.2 Spearman correlations of spinal mobility and mSASSS scores*

| | mSASSS | Lumbar part mSASSS | Cervical part mSASSS |
|---|---|---|---|
| Chest expansion | -0.49 | -0.59 | -0.46 |
| Finger to floor distance | 0.46 | -0.47 | 0.40 |
| Occiput to wall distance | 0.61 | 0.59 | 0.57 |
| Tragus to wall distance | 0.57 | 0.58 | 0.53 |
| Modified Schober | -0.60 | -0.64 | -0.40 |
| Lateral spinal flexion | -0.74 | -0.75 | -0.59 |
| Cervical rotation | -0.52 | -0.45 | -0.57 |
| Intermalleolar distance | -0.42 | -0.41 | -0.37 |
| BASMI | 0.76 | 0.75 | 0.61 |

mSASSS = modified Stoke Ankylosing Spondylitis Spinal Score, BASMI = Bath Ankylosing Spondylitis Metrology Index. * All correlations are statistically significant ($p < 0.05$).

To get a better insight in the effect of a different cut-off level for radiographic damage, we here discuss one example of the relation between radiographic damage and spinal mobility in detail. In figure 5.2 three scatter plots of the cumulative mSASSS versus modified Schober are represented. In each plot 2 lines are drawn. The vertical line indicates the cut-off level for the modified SASSS which was predefined, and set at 0, 3 and 6 mSASSS units respectively. From the figures it can be seen that according to a cut-off level of 0, 70% of the patients has radiographic damage, according to a cut-off level of 3 mSASSS units the prevalence of radiographic structural damage is 56%, and according to a cut-off level of 6 units the prevalence is 45%. The horizontal line indicates the cut-off level providing the highest accuracy for the modified Schober, (ROC analysis). Note that the optimal cut-off level for the modified Schober is dependent on the chosen cut-off level for radiographic damage. By drawing both cut-off levels in the scatter plots, four quadrants arise. In the right upper quadrant of figure 5.2a, all patients are represented which have an abnormal Schober test (≤2.3) and radiographic damage (mSASSS >0) (true positives: 33%). In the left lower quadrant of figure 2a, all patients are represented with a 'normal' (for this population) Schober test (>2.4) and no radiographic damage (mSASSS =0, true negatives: 28%). In the right lower quadrant are the patients with a normal modified Schober test but with radiographic damage (false negatives: 37%). In the right upper quadrant are the patients with an abnormal Schober test but without radiographic damage (false positives: 2%). These percentages can serve to calculate sensitivity (true positive rate) and specificity (true negative rate) of the spinal mobility assessment for discriminating between the absence and presence of radiographic damage. Figure 5.2 shows that if a higher cut-off level for radiographic damage is chosen, the percentage of true positives expectedly decreases (from 33% to 26%) whereas the percentage of true negatives increases (from 28% to 53%). The false positive rate remains approximately the same (2%–5%), but the percentage of false negatives becomes smaller (from 37% tot 19%) because of the increase of the true negatives. It is obvious that, even at the highest

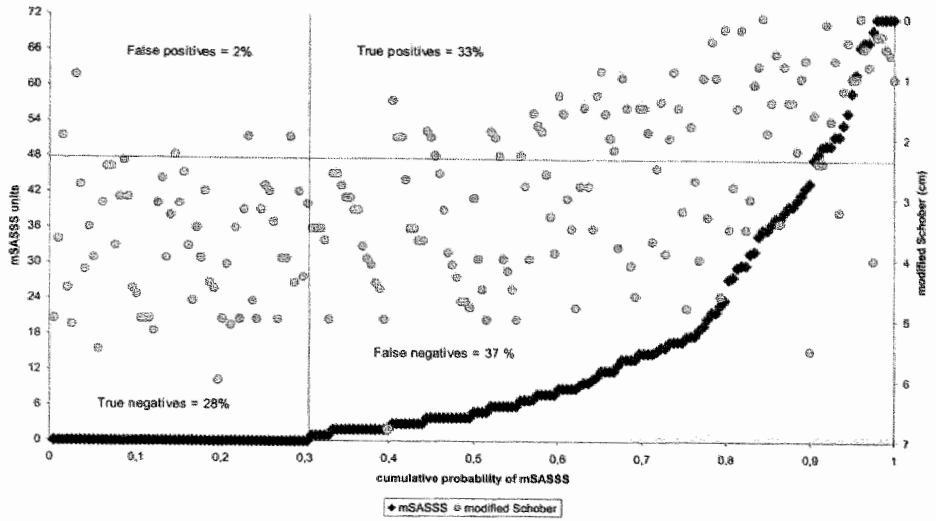
achievable level of accuracy, significant numbers of patients are falsely classified as having normal or abnormal X-rays of the spine, irrespective of the chosen cut-off level for radiographic damage.

For the modified Schober a LR+ of 5.7 and a LR- of 0.6 was found for a radiographic cut-off level of 0 (table 5.3). Since post-test probability relates to LR + (post-test odds = LR+ * pre-test odds), this means that finding an abnormal Schober's test (≤2.3 cm) increases the likelihood of radiographic damage from 0.70 to 0.93 (but is found in only 33% (true positives) of the patients). The value 0.93 is also called the positive predictive value (PPV+). A normal modified Schober test (>2.3 cm) decreases the likelihood of an abnormal X-ray of the spine from 0.70 to 0.43 (but is found in only 28% (true negatives) of the patients). The value 0.43 is called the negative predictive value (NPV-).

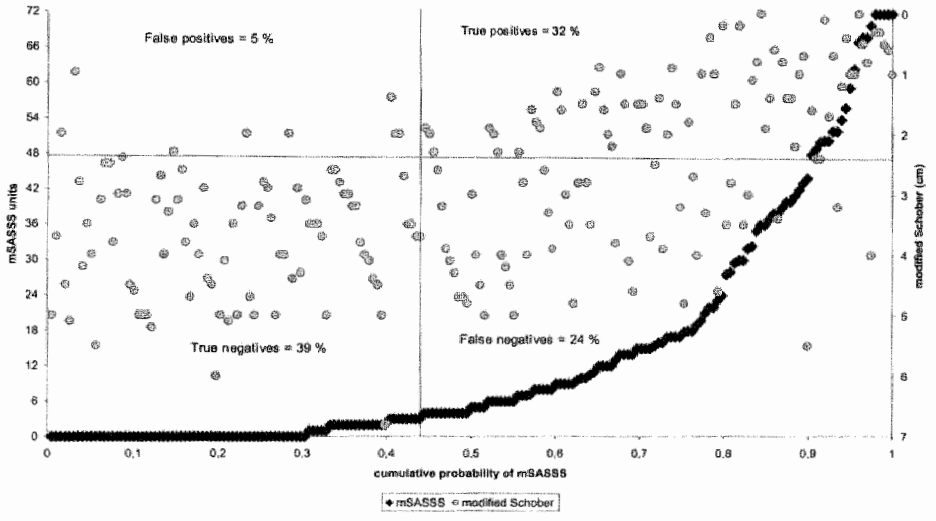
Calculating the PPV and NPV with the data from table 5.3 or percentages mentioned in figures 5.2b and 5.2c it can be seen that the PPV+ for modified Schober with a cut-off for mSASSS 3 = 0.88 (post test probability from 0.56 to 0.88) and the NPV- is 0.61, for a cut-off of mSASSS >6 the PPV+ =0.93 and NPV- =0.71.

Table 5.3 summarizes the results of the ROC-analysis and the determination of likelihood ratios for three different cut-offs of modified SASSS set as golden standard. For the golden standard defined at mSASSS >0 the AUC values range from 0.68 for intermalleolar distance to 0.85 for BASMI, which indicates that every spinal mobility measure to some extent differentiates between the absence and presence of radiographic damage. The corresponding levels of sensitivity and specificity, however, differed importantly between all spinal mobility measures: Sensitivity was highest for lateral spinal flexion 0.84 and lowest for Schober's index 0.47; specificity was highest for Schober's index 0.92 and lowest for cervical rotation 0.73. As a consequence of variability in sensitivity and specificity, LR+ and LR- importantly differed across spinal mobility measures. For mSASSS >0 set as golden standard the LR+ is highest for Schober's test (5.6) and lowest for cervical rotation(2.4); the LR- is lowest (which means best discriminatory) for lateral spinal flexion (0.2) and highest for Schober's index (0.6). Corresponding information can be found in table 5.3 for the two other cut-offs (mSASSS >3 and >6). Although the absolute values of the AUC, sensitivity and specificity differ with various cut-off levels, the relationship between the various spinal mobility measures remains constant. The BASMI composite index did not perform better than the lateral spinal flexion, irrespective of the chosen cut-off level for radiographic damage, but BASMI and lateral spinal flexion performed better than the other spinal mobility measures.

In table 5.4 it can be seen that for all investigated cut-off levels for radiographic damage a significant percentage of patients are misclassified.

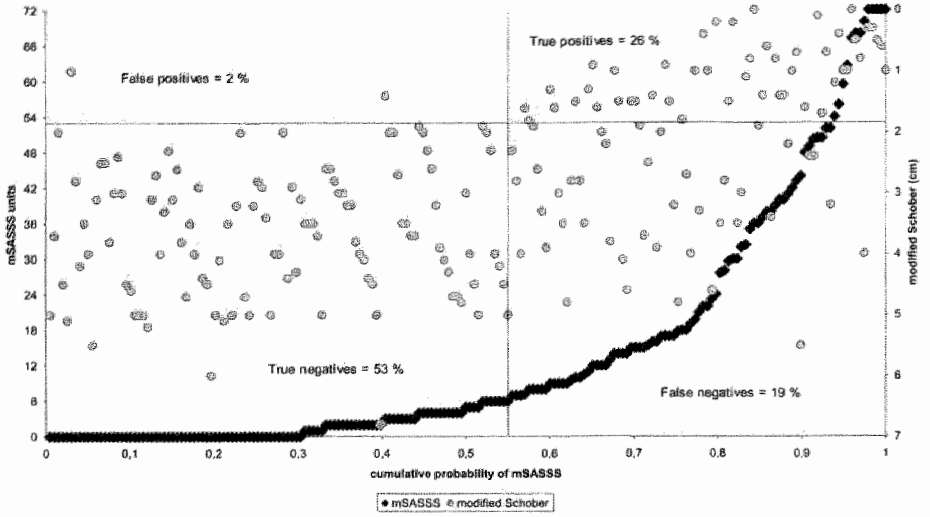



A: cut off mSASSS > 0




B: cut off mSASSS > 3

Figure 5.2 Scatter plots for cumulative mSASSS versus modified Schober with cut offs for mSASSS and modified Schober creating four quadrants.




C: cut off mSASSS > 6

Figure 5.2 Scatter plots for cumulative mSASSS versus modified Schober with cut offs for mSASSS and modified Schober creating four quadrants.

Table 5.3 Measurement properties of spinal mobility with radiographic damage set as 'golden standard'

| | Golden standard defined as mSASSS > 0 | | | | | | Golden standard defined as mSASSS > 3 | | | | | | Golden standard defined as mSASSS > 6 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *AUC* | *Cutoff* | *sens* | *spec* | *LR+* | *LR-* | *AUC* | *cutoff* | *sens* | *spec* | *LR+* | *LR-* | *AUC* | *cutoff* | *sens* | *spec* | *LR+* | *LR-* |
| Chest expansion (cm) | 0.75 | ≤4.9 | 0.68 | 0.75 | 2.7 | 0.4 | 0.79 | ≤ 4.2 | 0.69 | 0.81 | 3.7 | 0.4 | 0.84 | ≤ 4.1 | 0.75 | 0.82 | 4.1 | 0.4 |
| Finger to floor distance (cm) | 0.72 | >12.0 | 0.63 | 0.78 | 2.9 | 0.5 | 0.71 | >12.0 | 0.68 | 0.72 | 2.5 | 0.4 | 0.74 | > 11.8 | 0.78 | 0.71 | 2.7 | 0.5 |
| Occiput to wall distance (cm) | 0.71 | >1.5 | 0.59 | 0.80 | 2.9 | 0.5 | 0.77 | >2.4 | 0.68 | 0.84 | 4.2 | 0.4 | 0.81 | > 2.4 | 0.75 | 0.80 | 3.7 | 0.5 |
| Tragus to wall distance (cm) | 0.72 | >12.5 | 0.59 | 0.76 | 2.4 | 0.5 | 0.77 | >12.6 | 0.67 | 0.78 | 3.0 | 0.4 | 0.81 | > 14.1 | 0.62 | 0.86 | 4.5 | 0.5 |
| Modified Schober (cm) | 0.73 | ≤2.3 | 0.47 | 0.92 | 5.7 | 0.6 | 0.76 | ≤2.4 | 0.58 | 0.90 | 5.8 | 0.5 | 0.83 | ≤ 1.9 | 0.57 | 0.96 | 15.6 | 0.6 |
| Cervical rotation (degrees) | 0.71 | ≤68.0 | 0.63 | 0.73 | 2.4 | 0.5 | 0.72 | ≤54.0 | 0.46 | 0.92 | 5.7 | 0.6 | 0.75 | ≤ 54.0 | 0.51 | 0.89 | 4.6 | 0.5 |
| Lateral spinal flexion (cm) | 0.81 | ≤13.9 | 0.84 | 0.74 | 3.2 | 0.2 | 0.86 | ≤11.3 | 0.76 | 0.81 | 4.0 | 0.3 | 0.88 | ≤ 9.4 | 0.74 | 0.86 | 5.2 | 0.2 |
| Intermalleolar distance (cm) | 0.68 | ≤102.0 | 0.52 | 0.83 | 3.0 | 0.6 | 0.65 | ≤102.0 | 0.53 | 0.73 | 2.0 | 0.7 | 0.71 | ≤ 104.0 | 0.63 | 0.70 | 2.1 | 0.6 |
| BASMI | 0.85 | >3.0 | 0.75 | 0.84 | 4.8 | 0.3 | 0.86 | >3.2 | 0.78 | 0.79 | 3.7 | 0.3 | 0.89 | > 3.6 | 0.80 | 0.84 | 5.1 | 0.3 |

BASMI = Bath Ankylosing Spondylitis Metrology Index; cut off = value with the optimal sensitivity and specificity according to a Receiver Operator Characteristic (ROC) analysis; AUC = area under the curve; sens = sensitivity; spec = specificity; LR + = likelihood ratio of a positive test; likelihood ratio of a negative test.

Table 5.4 Percentages true positives and negatives, and false positives and negatives for measurement properties of spinal mobility, with radiographic damage set as 'golden standard'

| | Golden standard defined as mSASSS > 0 | | | | Golden standard defined as mSASSS > 3 | | | | Golden standard defined as mSASSS > 6 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *true positive* | *true negative* | *false positive* | *false negative* | *true positive* | *true negative* | *false positive* | *false negative* | *true positive* | *true negative* | *false positive* | *false negative* |
| Chest expansion (cm) | 48 | 22 | 8 | 22 | 39 | 35 | 8 | 18 | 34 | 45 | 10 | 11 |
| Finger to floor distance (cm) | 44 | 24 | 6 | 26 | 38 | 31 | 12 | 18 | 35 | 39 | 10 | 16 |
| Occiput to wall distance (cm) | 41 | 24 | 12 | 29 | 38 | 37 | 7 | 18 | 34 | 44 | 11 | 11 |
| Tragus to wall distance (cm) | 41 | 23 | 7 | 29 | 38 | 33 | 10 | 19 | 28 | 47 | 8 | 17 |
| Modified Schober (cm) | 33 | 28 | 2 | 37 | 32 | 39 | 5 | 24 | 26 | 53 | 2 | 19 |
| Cervical rotation (degrees) | 44 | 22 | 8 | 26 | 26 | 40 | 3 | 31 | 23 | 49 | 6 | 22 |
| Lateral spinal flexion (cm) | 60 | 21 | 8 | 11 | 44 | 34 | 8 | 14 | 35 | 45 | 8 | 12 |
| Intermalleolar distance (cm) | 36 | 25 | 5 | 34 | 30 | 32 | 12 | 27 | 28 | 39 | 16 | 17 |
| BASMI | 53 | 24 | 5 | 18 | 44 | 34 | 9 | 13 | 37 | 45 | 9 | 9 |

BASMI = Bath Ankylosing Spondylitis Metrology Index .

# Discussion

This study demonstrates that assessing spinal mobility can not be a proxy for the assessment of spinal structural damage assessed by radiographs. Even the spinal mobility measure with the highest level of accuracy still misclassifies a significant number of patients as having or not having radiographic damage. This observation does not dispute the concept that radiographic damage is associated with a decreased spinal mobility. Significant correlations were found for mSASSS scores and spinal mobility measures on a group level (highest correlation for lateral spinal flexion and BASMI). However these correlation coefficients relate to the entire group, and are disproportionally influenced by small numbers of observations with both high structural damage scores and strongly impaired spinal mobility, as shown by the probability plots. The probability plots show unequivocally that on the individual patient level the association between spinal mobility measures and radiographic damage can be very variable.
Of all spinal mobility measures the intermalleolar distance had the lowest correlation with radiographic damage. An explanation for this might be that intermalleolar distance assesses the mobility in the hip joint whereas the mSASSS only assessed the spine. Correlation coefficients also disguise the phenomenon of variability in spinal mobility in the absence of structural damage. And at last, high correlation coefficients do not necessarily indicate a high level of discriminatory power in the individual patient level, as was shown here. Rephrased differently, not every patient with radiographic structural damage has a reduced spinal mobility, and not every patient with reduced spinal mobility has radiographic damage. This finding implicates that both the assessment of spinal mobility and radiographic assessment have an additive place in the outcome measurement of AS.

The considerable number of patients falsely classified as having or not having radiographic damage deserves some further explanation. Patients without radiographic damage but with impaired spinal mobility (false positives) may have involvement of other structures that are not visualized by means of radiography (inflammation of soft tissue for example), or structural damage (e.g. in thoracic spine or zygapophyseal joints) not captured with the mSASSS. Impairment of spinal mobility due to inflammation is underscored by the observation in clinical trials that TNF-blocking therapy may increase spinal mobility after only a few months of treatment. Another limiting factor may be that the modified SASSS only takes structural damage in the anterior site of the lumbar and cervical spine into account. Exclusion of the thoracic spine and of the posterior site of the spine may result in an underestimation of true structural damage which causes limitation of spinal mobility but is not picked up by the mSASSS. Also exclusion of the facet joints, which play a major role in spinal mobility, may result in an underestimation.

DeVlam et al found a relationship between involvement of the facet joints and the presence of syndesmophytes, suggesting that the facet joint is primarily involved[12]. Other structures that are not incorporated in the mSASSS are the anterior and posterior ligaments. If a ligament shows ossification, but no structural damage of the vertebrae is seen, then a normal mSASSS score will be assigned to this patient, while the patient may experience severe limitation of spinal mobility. So, the mSASSS is certainly not a perfect score to represent all possible radiographic abnormalities. The Bath Ankylosing Radiography Index (BASRI) is a method that incorporates much more abnormalities, but the BASRI is a rather rough method, and correlation coefficients for BASRI scores and measures of spinal mobility were not better than those found in this study[8].
Magnetic resonance imaging (MRI) can visualize both soft tissue and bone. Further, MRI can visualize inflammatory activity, as well as the chronic irreversible damage that are typically seen on X-rays of the spine. Since involvement of soft tissue may be an important factor determining spinal mobility, it is of interest to investigate the correlation between both inflammatory activity (soft tissue) and structural damage (bone) assessed by MRI, and spinal mobility in future research.

Part of the explanation of normal spinal mobility in the presence of significant radiographic damage ("false negatives") is formed by the choice of the cut-off point for radiographic progression. It was seen that with a higher cut-off for radiographic damage the cut-off for the spinal mobility measure indicated a worse value. It is very well conceivable that a patient with a minimal mSASSS score (for example an mSASSS of 2 which indicates sclerosis, or erosions or squaring on two vertebrae or one syndesmofyte and no involvement of other structures does not experience a limited spinal mobility.
There is also a number of patients with high mSASSS scores, indicating that at least part of the spine has severe structural damage, that still have a good spinal mobility. This group of patients is small but existing. Apparently, patients can compensate impaired mobility by severe structural damage in parts of the spine to a certain degree. It is obvious, however, that the highest levels of radiographic damage (patients at the ceiling of the scoring range) all experience severely impaired mobility, which adds to the validity of the construct that radiographic damage compromises spinal mobility.
We investigated whether the results that we found with regard to the discriminative power of several spinal mobility measures were sensitive to the predefined cut-off level for radiographic damage. This question is relevant because many investigators in this field will dispute the reliability of a cut-off level for radiographic damage equal to zero. Since scoring radiographic damage is prone to all sources of measurement error and biases that may track structural damage score towards higher levels, a cut-off level above zero could be considered more realistic. However, although we found several small differences in the performance of spinal mobility measures, the general picture was similar, irrespective of the chosen cut-off level for radiographic damage.

We think that the arbitrary choice of cut-off levels of mSASSS >3 and mSASSS >6 appeared to be a good choice. Setting the cut-off level higher than mSASSS >6 was not appropriate since the median of the mSASSS scores was at 5.0, therefore a cut-off much higher than the median would omit the majority of patients, 55% of patients had a score of mSASSS of 6 units or lower.

A theoretical limitation of this study may be that the results are only valid within the OASIS cohort. We do not believe that external validity is jeopardized here, since OASIS includes unselected, consecutive AS patients, and we have shown here that the entire range of spinal mobility impairment and radiographic damage is actually included in the cohort.

# References


1. Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. Ankylosing spondylitis: the correlation between a new metrology score and radiology. Br J Rheumatol 1995;34:767-70.
2. Viitanen JV, Kokko ML, Lehtinen K, Suni J, Kautiainen H. Correlation between mobility restrictions and radiologic damage in ankylosing spondylitis. Spine 1995;20:492-6.
3. Viitanen JV, Kokko ML, Heikkila S, Kautiainen H. Neck mobility assessment in ankylosing spondylitis: a clinical study of nine measurements including new tape methods for cervical rotation and lateral flexion. Br J Rheumatol 1998;37:377-81.
4. Viitanen JV, Heikkila S, Kokko ML, Kautiainen H.Clinical assessment of spinal mobility measurements in ankylosing spondylitis: a compact set for follow-up and trials? Clin Rheumatol 2000;19:131-7.
5. Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van der Tempel H, van der Linden S. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. J Rheumatol 1999;26: 980-4.
6. Creemers M, Franssen M, van 't Hof M, Gribnau F, van de Putte L, van Riel P, Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. Ann Rheum Dis 2004. Published Online First [March 29,2004]
7. Spoorenberg A, de Vlam K, van der Linden S, Dougados M, Mielants H, van der Tempel H, van der Heijde D. Radiological scoring methods in ankylosing spondylitis: reliability and sensitivity to change over one year. J Rheumatol 1999;26:997-1002.
8. Wanders AJB, Landewe RBM, Spoorenberg A, Dougados M, van der Linden S, Mielants H, van der Tempel H, van der Heijde DM. What is the most appropriate radiologic scoring method in Ankylosing spondylitis clinical trials. A comparison based on the OMERACT filter. Arthritis Rheum 2004;50:2622-32
9. Jenkinson TR, Mallorie PA, Whitelock HC, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis (AS). The Bath AS Metrology Index. J Rheumatol 1994;21:1694-8.
10. Jones SD, Porter J, Garrett SL, Kennedy LG, Whitelock H, Calin A. A new scoring system for the Bath Ankylosing Spondylitis Metrology Index (BASMI). J Rheumatol 1995;22:1609.
11. Landewe RBM, van der Heijde DMFM, Principles of assessment from a clinical perspective. Best Pract Res Clin Rheumatol 2003;17:365-79.
12. de Vlam K, Mielants H, Veys EM, Involvement of the zygapophyseal joint in ankylosing spondylitis: relation to the bridging syndesmophyte. J Rheumatol 1999;26:1738-45.

# Chapter 6

# How the type of risk reduction influences required sample sizes in randomized clinical trials

K Bruynesteyn, A Wanders, R Landewé, D van der Heijde

Ann Rheum Dis 2004;63:1368-71

# Abstract

In order to increase change (and treatment contrast) between groups, randomized clinical trials (RCT) often include patients with a high risk on a particular outcome, by inclusion criteria that select on predictors for that outcome. The general view is that such a strategy increases the statistical power, and thus limits the number of patients required for that RCT.
How the selection of patients influences the power and thus the sample size required, depends on how an intervention reduces the individual risk: by an absolute or relative risk reduction model (ARR or RRR-model, respectively). We here explain both models and show that if a treatment mainly acts according to the RRR-model, selection of patients with a high prior risk will result in higher statistical power, and thus smaller sample sizes. If a treatment acts mainly according to the ARR-model, selection of patients by prior risk may have divergent effects on the sample size required in trials with dichotomous outcome measures. Statistical power is worst in case of a prior risk of on average ~50%, which can be the case when including a mixture of high, intermediate and low risk patients. As a consequence, the required sample size will be higher.
To explore the relationship between prior risk and treatment effect in rheumatological trials, we analyzed the data of two trials in which disease modifying anti-rheumatic drugs (DMARDs) in rheumatoid arthritis (RA) were investigated. Radiological progression (above or below the median) was defined as outcome, and joint damage at baseline (present or absent) as prior risk. In both trials, the DMARD treatment with highest efficacy seemed to act mainly according to the ARR-model.
In order to design appropriately powered RCTs with dichotomous outcome measures, it is relevant to be informed about whether tested therapies work mainly according to a RRR or according to an ARR- model. If data are lacking, it is best to select patients with a high prior risk on the outcome that should be prevented by the therapy, under the assumption that this is feasible.

In order to increase the contrast between an intervention and a control group, randomized clinical trials (RCTs) often include patients with a high prior risk on the outcome of interest by selecting on baseline predictors for that outcome. The general view is that such a strategy limits the number of patients required, by increasing the statistical power. In RCTs in RA, one important outcome is progression of radiological damage. It is known that joint damage present at baseline (prior risk) is one of the strongest predictors for further progression of radiological joint damage[1-4]. Therefore, we expected fewer patients would be needed for a RCT if patients were selected for the presence of joint damage at baseline. Somewhat surprisingly, power calculations learned that this expectation was false. To elucidate this counterintuitive observation, we performed a literature search, and we found that statistical power is determined by a critical relationship between the prior risk on a particular dichotomous outcome (this is the baseline risk, independent of treatment), and on how a particular treatment exerts its efficacy with respect to this prior risk. We will explain this relationship step-by-step, and illustrate it with hypothetical and authentic data.

The risk of an individual patient on a particular outcome can be reduced in two different ways, according to the following two models: the absolute risk reduction model, and the relative risk reduction model. For understanding these models it is necessary to be familiar with the terms absolute risk reduction (ARR) and relative risk reduction (RRR).
Consider a RCT comparing an active treatment with placebo. The outcome of interest is binomial (*event* or *no event*) and the active treatment can reduce the probability of that event. So, the event rate in the treatment group ($P_t$) is supposed to be lower than the event rate in the placebo group ($P_p$). Absolute risk reduction (ARR) is the difference in the event rates between the placebo and the treatment group ($P_p$-$P_t$). Relative risk is defined here as the ratio of the event rate in the treatment group and that in the placebo group ($P_t$/ $P_p$). RRR is defined as the reduction of the event rate in the treatment group, in proportion to that in the placebo group, or: ($P_p$ - $P_t$)/ $P_p$. This is mathematically similar to (1 – the relative risk), or: (1 – $P_t$/ $P_p$). RRR is usually expressed as a percentage, or: (1-($P_t$/ $P_p$) * 100%).
If a therapy works according to a RRR-model, the RRR remains constant, irrespective of the prior risk, whereas the ARR varies with the prior risk. If a therapy works according to an ARR-model, the ARR remains constant irrespective of prior risk and then the RRR varies with different prior risks.
Suppose that we distinguish three subgroups, defined by the prior risk on a particular outcome: one group with a low prior risk, one group with an intermediate prior risk, and one group with a high prior risk. Note that prior risk does not directly refer to the event rate in the control group, but rather to patient characteristics (such as age, gender, disease aetiology, concomitant conditions or disease status) measured at baseline, and known for their ability to influence the probability of the negative outcome. Suppose also that without any (adequate) treatment the negative outcome will occur in 20%, 60% and 90% of the patients per group, respectively. Assume that therapy works purely

according to the RRR-model (table 6.1), providing a RRR of 50%. So, treatment will lower the event rate from 20% to 10% (50% reduction) in the low risk group, from 60% to 30% (50% reduction) in the intermediate risk group, and from 90% to 45% (50% reduction) in the high-risk group. The ARR is now 10%, 30%, and 45% respectively. These figures can be used to calculate estimated sample sizes for future clinical trials. We calculated the samples sizes for two-sided statistical testing using the power calculator of the UCLA department of statistics with the 2-sample arcsine approximation of the binomial distribution, with alpha set at 0.05 and beta at 0.20[5]. From table 6.1A (sample sizes per group) it is obvious that the baseline risk importantly influences the appropriate sample size. In other words: In this scenario, selecting high-risk patients will increase the statistical power of a trial.

Table 6.1 Example of the absolute and relative risk model illustrated with data of clinical trials

A Example of the relative risk reduction model

| | Control | Treatment | RRR | ARR | Sample size per group |
|---|---|---|---|---|---|
| Low baseline risk | 20% | 10% | 50% | 10% | 195 |
| Medium baseline risk | 60% | 30% | 50% | 30% | 41 |
| High baseline risk | 90% | 45% | 50% | 45% | 14 |

B Example of the absolute risk reduction model

| | Control | Treatment | RRR | ARR | Sample size per group |
|---|---|---|---|---|---|
| Low baseline risk | 20% | 10% | 50% | 10% | 195 |
| Medium baseline risk | 60% | 50% | 17% | 10% | 387 |
| High baseline risk | 90% | 80% | 11% | 10% | 195 |

C CSA + MTX compared with CSA alone

| | CSA + placebo** | CSA + MTX** | RRR | ARR | Sample size per group |
|---|---|---|---|---|---|
| Low baseline risk* | 50%(n=28) | 25%(n=32) | 50% | 25% | 57 |
| High baseline risk | 79%(n=29) | 50%(n=26) | 37% | 29% | 40 |

D COBRA treatment compared with SSZ alone

| | SSZ** | COBRA** | RRR | ARR | Sample size per group |
|---|---|---|---|---|---|
| Low baseline risk* | 40% (n=30) | 18%(n=38) | 54% | 22% | 68 |
| High baseline risk | 82% (n=34) | 58%(n=33) | 30% | 25% | 51 |

RRR = [1 - (outcome rate treatment group divided by outcome rate controls) * 100]; ARR = [outcome rate control group – outcome rate therapy group]; Sample size calculation based on alpha=0.05, beta=0.20, two sided. CSA = cyclosporine, MTX = methotrexate, SSZ = sulphasalazine, COBRA treatment = step-down prednisolone + MTX + SSZ; * Risk groups based on baseline radiographic damage (below or above the median); ** Percentage of patients with radiographic progression larger than the median.

Now assume that a therapy works purely according to the ARR-model (table 6.1B), providing an ARR of 10%. Based on the same prior risk percentages, treatment will lower the event rate from 20% to 10% (10% reduction) in the low risk group, from 60% to 50% (10% reduction) in the intermediate risk group, and from 90% to 80 % (10% reduction) in the high-risk group. The RRR is now 50% ((1– (10% / 20%)) * 100) in the low risk group, 17% ((1 – (50% / 60%)) * 100) in the intermediate risk group and 11% ((1 – (80% / 90%)) * 100) in the high-risk group, respectively. If again sample sizes are estimated using these figures, it becomes clear that the same sample sizes that are required in low and high-risk groups to statistically demonstrate an ARR of 10%. But to demonstrate an ARR of 10% for the intermediate risk group, a much larger sample size is needed. This can be explained by the fact that for intermediate risk groups the probability on the occurrence of a negative outcome is almost similar to the probability on the occurrence of a positive outcome. Note that an intermediate risk group may either include patients with an intermediate prior risk only, or a mix of patients with a low, high and intermediate risk.
Summarized: depending on how a therapy reduces the individual's risk on a particular negative outcome (according to an ARR – or a RRR-model), the patient selection for the trial with respect to the prior risk may influence the statistical power, and accordingly the sample size required for that trial. If a therapy (mainly) acts according to the RRR-model, trials including patients with prognostic variables for a negative outcome (high prior risk) will yield more statistical power, and require a lower sample size. If a therapy acts mainly according to the ARR- model, selection of patients with respect to prognostic variables has a different effect on sample size: From a statistical point of view, it would be wise to avoid groups of patients with an average prior risk on the negative outcome of approximately 50%. Trials with such patient groups provide less statistical power, as compared to trials with patient groups with a low or a high prior risk.
In the sparse literature that evaluated the relationship between treatment effect and the prior risk, it is suggested that the RRR is constant across the usual spectrum of prior risks in the vast majority of treatments (i.e. following the RRR-model)[6,7]. Nevertheless exceptions have been described. A large study evaluating the stroke risk in patients treated with aspirin, showed a decreasing RRR by increasing prior risk[8]. However, the ARR also decreased by increasing baseline risk, so actually a mix of both models seemed to be operative. Obviously, treatments do not always act strictly according the RRR- or to the ARR- model: Mixed models can also be found.

To our best knowledge, in rheumatology no research has been performed to examine the relationship between prior risks and models of risk reduction, with respect to treatment effect. Recently, radiographic data of a RCT comparing methotrexate (MTX) + cyclosporine (CsA) versus CsA monotherapy have become available[10]. We used these data to investigate this relationship. Treatment effect was defined here as the reduction in the proportion of patients with the negative outcome, being radiographic progression $\geq$ the median group level, at one year. The prior risk on radiographic progression

(baseline risk) was based on the radiological damage at baseline, above or below the median. Table 6.1C shows that the RRR was not similar in the two baseline risk groups (50% in the low risk group *vs.* 37% in the high-risk group): The addition of MTX to CsA appeared not to follow a pure RRR model, but rather an ARR model (the ARR was approximately similar in both risk groups (25% *vs.* 29%)).

We further explored the radiographic data of the COBRA trial (table 6.1D), in order to confirm the phenomenon of a decreasing RRR by an increasing baseline risk. In the COBRA trial[9], combination treatment with step-down prednisolone, MTX and sulfasalazine (SSZ) was compared with SSZ monotherapy. We again defined treatment effect as the reduction in patients with radiographic progression above the median level in one year.

Table 6.1D shows that the ARR (combination therapy, compared with SSZ alone) was almost similar in the two risk groups (22% *vs.* 25%), and that the RRR decreased by increasing baseline risk (54% *vs.* 30%).

We statistically compared the treatment effects in both subgroups of baseline damage, in order to test the null hypothesis that the relative risk reductions were similar (test of interaction, as recommended by Matthews & Altman, and Altman & Bland[11,12]. In neither of both studies, a significant interaction could be demonstrated (P=0.29 for the difference in treatment effects in the MTX + CsA study, and P=0.17 for the difference in treatment effect in the COBRA study). For the test of interaction, as well as for an interpretation and discussion of the lack of statistical interaction, we refer to Appendix I.

These observations can of course not be generalized to the effect of all DMARDS on radiological joint damage or on other outcome measures. However, if it is true that some of the DMARDS decrease radiological progression by means of an ARR rather than a RRR model, selection of patients with an average prior risk on radiological progression of about 50% should be avoided when a RCT with such DMARD is assigned (table 6.1B). But only a proportion of all patients with RA that eventually show radiological progression will have radiological joint damage at inclusion. This makes it difficult to selectively enroll patients with a high baseline risk, and the actual patient accrual may include patients with intermediate, rather than with high baseline risk.

We want to make a few additional remarks. First, a trialist who is developing a trial has to make a choice between aiming at a mixture of high-, intermediate- and low-risk patients, and focusing on just one category. For generalisability purposes one may choose to include all types of risk patients. However, we showed here that this might lead to larger sample sizes. On the other hand, one should consider whether the preferred inclusion of high-risk patients is feasible. If high-risk patients are difficult to include for any reason, the argument of an appropriate recruitment rate may outweigh the argument of limited sample sizes by the selective inclusion of high-risk patients.

Patient selection in RCTs is often based on characteristics that are predictive of a negative outcome. Aim of this report was partly to show that statistical power is dependent on the level of that prior risk, as well as on how treatment actually reduces

that risk. This is a different approach as compared to selecting patients on the individual likelihood of *responding* to a particular treatment. Selecting patients with a high probability of responding to a particular treatment will also result in a larger treatment effect and thereby an increasing ARR and RRR. This means more statistical power, and as a consequence, a smaller required sample size. Increased responsiveness at a group level in RCTs can be promoted in two different ways[13]. The first is by taking measures that result in a general increase of patient's compliance. Those who take their medicine (appropriately) might respond better than those who do not take their medicine (appropriately). The second way to promote responsiveness is by identifying subgroups of patients that are intrinsically responsive to the particular treatment. But patient characteristics that are predictive of a high response to any treatment have hardly be identified in rheumatology up to now. This will be an important research area in the future.

Our final remark refers to the models used to describe the relationship between the prior risk and the treatment effect. These models are based on discrete binomial outcome measures. The primary outcome measure of a trial can also be a continuous variable. Although the models cannot simply be translated to continuous measures, there are no arguments why treatment effects should act differently when measured on a continuous scale as compared to a dichotomous one.

We here conclude that patient selection with regard to factors predictive of the outcome that should be influenced by therapy has an impact on the statistical power of RCTs with dichotomous outcome measures. The precise direction of that impact depends not only on the level of prior risk, but also on whether the ARR or the RRR remains constant irrespective of the prior risk. As a rule of thumb, statistical power is best guaranteed by selecting high-risk patients, because this scenario omits the dependency of the type of risk reduction. Better insight in prediction of individual responsiveness may further increase statistical power and decrease required sample size.

# Appendix I

The statistical proof of whether an absolute- or a relative risk reduction model is operative is difficult and circumstantial. The statistical inference refers to the comparison of treatment effects across subgroups of a randomised clinical trial. The null hypothesis is that the difference of relative risks in both subgroups is zero (the so-called test of interaction).

Treatment effects are represented here as relative risks (*e.g.*, the risk of having radiographic progression above the median when treated with CsA + MTX in relation to that risk when treated with CsA + placebo). The two subgroups are the low baseline risk group and the high baseline risk group.

In an absolute risk reduction model, the absolute risk reductions are similar in both subgroups, whereas the relative risk reductions in both subgroups are different. The latter effect refers to interaction, and can be tested statistically, as shown below. The equivalence of absolute risk reductions is difficult to prove. Theoretically, the proof of different relative risk reductions does not suffice, since this does not prove that absolute risk reductions in both subgroups are similar.

Here we show the test of interaction performed in both clinical trials that we have used to corroborate the argument of the absolute risk reduction model in our article. Table 6.2A. and table 6.2B. show the raw data of both trials, that are also shown in the article, but now tabulated in a different order, in order to improve readability with respect to the inferences below. Note that we have calculated the relative risk here as a starting point for the inferences. We also added, as an illustration, the absolute and relative risk reductions calculated in the article. Table 6.3. shows the statistical inferences necessary to test for interaction in both trials. The inference is complicated because of the logarithmic transformation (and the necessary back-transformation). The test of interaction is not statistically significant in neither of both trials, indicating that the null hypothesis that both relative risk reductions are similar cannot be rejected (*or.* it is not proven that the relative risk reductions in both subgroups are different). Since the subgroup sizes are small (lack of statistical power), since a test of interaction is quite conservative, and since the difference of relative risk reductions is considerably higher than the difference of absolute risk reductions in both trials, the failure to prove interaction here does not disqualify our statement that an absolute, rather than a relative risk reduction model is operative here. Note with respect to Table 6.2, line 10, that the ratio of relative risks (sometimes abbreviated as RRR) is not similar to the relative risk reduction in this article (abbreviated by us as RRR).

Table 6.2 The test of interaction performed in the CsA + MTX study, and the COBRA study.

A CsA + MTX study:

| | Low baseline risk | | High baseline risk | |
|---|---|---|---|---|
| *Progression:* | CsA+MTX | CsA+placebo | CsA+MTX | CsA+placebo |
| Yes | 8 (A) | 14 (B) | 13 (A) | 23 (B) |
| No | 24 (C) | 14 (D) | 13 (C) | 6 (D) |
| Total | 32 (A+C) | 28 (B+D) | 26 (A+C) | 29 (B+D) |
| Relative risk (RR) for CsA+MTX | = (8/32) : (14/28) | | = (13/26) : (23/29) | |
| vs. CsA+placebo | = 0.50 | | = 0.63 | |
| Absolute risk reduction (ARR) | = (14/28) – (8/32) | | = (23/29) – (13/26) | |
| | = 0.25 | | = 0.29 | |
| Relative risk reduction (RRR) | = (1 – (8/32 : 14/28)) | | = (1 – (13/26) : (23/29)) | |
| | = 0.50 | | = 0.37 | |

B COBRA study:

| | Low baseline risk | | High baseline risk | |
|---|---|---|---|---|
| *Progression:* | SSZ | COBRA | SSZ | COBRA |
| Yes | 12 (A) | 7 (B) | 28 (A) | 19 (B) |
| No | 18 (C) | 31 (D) | 6 (C) | 14 (D) |
| Total | 30 (A+C) | 38 (B+D) | 34 (A+C) | 33 (B+D) |
| Relative risk (RR) for COBRA vs. | = (7/38) : (12/30) | | = (19/33) : (28/34) | |
| SSZ | = 0.46 | | = 0.70 | |
| Absolute risk reduction (ARR) | = (12/30) – (7/38) | | = (28/34) – (19/33) | |
| | = 0.22 | | = 0.25 | |
| Relative risk reduction (RRR) | = (1 – (7/38 : 12/30)) | | = (1 – (19/33) : (28/34)) | |
| | = 0.54 | | = 0.30 | |

Table 6.3 Comparing relative risks in two risk groups

| | COBRA trial | | CSA+MTX trial | |
|---|---|---|---|---|
| | Low baseline risk | High baseline risk | Low baseline risk | High baseline risk |
| 1. RR | 0.46 | 0.70 | 0.50 | 0.63 |
| 2. log(RR)* | -0.7765 | -0.3567 | -0.6931 | -0.4620 |
| 3. SE (log(RR))** | 0.4081 | 0.1692 | 0.3598 | 0.2178 |
| *Difference between the logarithms of relative risks:* | | | | |
| 4. $d(=E_1 - E_2)$ | -0.4198 | | -0.2311 | |
| 5. SE ($d$) | = $\sqrt{(0.4081^2 + 0.1692^2)}$ = 0.4419 | | 0.4206 | |
| 6. CI ($d$) | = -0.4198 ± 1.96 x 0.4419, or: -1.2859 to 0.4463 | | -1.0555 to 0.5933 | |
| 7. Test of interaction | $z$ = -0.4198 / 0.4419 = 0.95 (P=0.17) | | $z$ = -0.55 (P=0.29) | |
| *Ratio of relative risks (ratio):* | | | | |
| 8. Ratio=exp($d$) | =exp(-0.4198) = 0.66 | | 0.79 | |
| 9. CI (ratio) | =exp(-1.2859) to exp(0.4463), or: 0.28 to 1.56 | | 0.35 to 1.81 | |

* Values obtained by taking the natural logarithm; ** The SE (log (RR)) can be calculated as follows: SE (log (RR)) = $\sqrt{(1/A - 1/(A+C) + 1/B - 1/(B+D))}$, in which A, B, C and D refer to the raw numbers in the 2X2 tables (table 6.2A and table 6.2B)

# References


1. Uhlig T, Smedstad LM, Vaglum P, Moum T, Gerard N, Kvien TK. The course of rheumatoid arthritis and predictors of psychological, physical and radiographic outcome after 5 years of follow-up. Rheumatology (Oxford) 2000;39:732-741.
2. Boers M, Kostense PJ, Verhoeven AC, van der Linden S. Inflammation and damage in an individual joint predict further damage in that joint in patients with early rheumatoid arthritis. Arthritis Rheum 2001;44:2242-6.
3. Combe B, Dougados M, Goupille P, Cantagrel A, Eliaou JF, Sibilia J, Meyer O, Sany J, Daures JP, Dubois A. Prognostic factors for radiographic damage in early rheumatoid arthritis: a multiparameter prospective study. Arthritis Rheum 2001;44:1736-43.
4. Drossaers-Bakker KW, Zwinderman AH, Vlieland TP, Van Zeben D, Vos K, Breedveld FC, Hazes JM. Long-term outcome in rheumatoid arthritis: a simple algorithm of baseline parameters can predict radiographic damage, disability, and disease course at 12-year followup. Arthritis Rheum 2002;47:383-90.
5. http://calculators.stat.ucla.edu/powercalc/
6. Osiri M, Suarez-Almazor ME, Wells GA, Robinson V, Tugwell P. Number needed to treat (NNT): implication in rheumatology clinical practice. Ann Rheum Dis 2003;62:316-21
7. McAlister FA. Commentary: relative treatment effects are consistent across the spectrum of underlying risks...usually. Int J Epidemiol 2002;31:76-7.
8. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995;345:1616-9.
9. Boers M, Verhoeven AC, Markusse HM, van de Laar M, Westhovens R, van Denderen JC, van Zeben D, Dijkmans BAC, Peeters AJ, Jacobs P, van den Brink HR, Schouten HJA, van der Heijde D, Boonen A, van der Linden S. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309-18.
10. Gerards AH, Landewé RBM, Prins APA, Bruijn GAW, Goei Thè HS, Laan RFJM, Dijkmans BAC. Cyclosporin A monotherapy versus cyclosporin A and methotrexate combination therapy in patients with early rheumatoid arthritis: a double blind randomised placebo controlled trial. Ann Rheum Dis 2003;62:291-96.
11. Matthews JN, Altman DG. Statistical notes: Interaction 3: How to examine heterogeneity. BMJ 1996;313:862
12. Altman DG, Bland JM. Statistical notes: Interaction revisited: the difference between two estimates. BMJ 2003;326:219
13. Sackett DL. Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). CMAJ 2001;165:1226-37.

# Chapter 7

# Non-steroidal anti-inflammatory drugs inhibit radiographic progression in patients with ankylosing spondylitis: A randomised clinical trial

A Wanders, D van der Heijde, R Landewé, J-M Béhier, A Calin, I Olivieri, H Zeidler, M Dougados

*Arthritis Rheum 2005: accepted for publication*

# Abstract

Introduction
A 2-year randomized controlled trial was performed to test the hypothesis that long-term continuous non steroidal anti inflammatory drugs (NSAID)-use, in comparison with NSAID-use on demand only, influences radiographic progression in ankylosing spondylitis (AS).

Methods
Patients with AS (n=215), who had previously participated in a 6-week randomized double blind clinical trial, which compared celecoxib, ketoprofen and placebo, were randomly allocated to receive either a continuous treatment or an on demand treatment in case of symptoms. for a period of two years. All patients started on celecoxib 100 mg twice daily, which patients could increase to 200 mg twice daily or switch to another NSAID, but keeping the same treatment strategy. Structural changes were assessed by radiographs of the lumbar and cervical spine, and scored according to the modified Stoke Ankylosing Spondylitis Spinal Score (SASSS) by one observer blinded for treatment strategy and time order of the radiographs. Statistical analyses included a between group comparison of 1) radiographic progression scores (by Mann-Whitney U-test); 2) time-averaged values of variables reflecting signs and symptoms of AS (by linear regression analysis) and 3) the frequency of reported site-specific adverse events (by Chi-square test, or Fisher's exact test if appropriate).

Results
Complete sets of radiographs were available for 76 of the 111 patients in the continuous treatment group and for 74 of the 104 patients in the on demand group. The mean radiological progression was 0.4 (SD 1.7) in the continuous treatment group and 1.5 (SD 2.5) in the on demand treatment group (p=0.002 for the between-group difference). Parameters reflecting signs and symptoms were not statistically significantly different between both groups. The between-group difference in radiographic progression did neither disappear after adjustment for baseline values of radiographic damage or disease activity variables and for time-averaged values of disease activity variables, nor after imputation of missing data. Relevant adverse events, such as hypertension (in 9% vs. 3%), abdominal pain (in 11% vs. 6%) and dyspepsia (in 41% vs. 38%), tended to occur more frequently in the continuous group, but the differences were not statistically significant.

Conclusion
A strategy of continuous use of NSAIDs reduces radiographic progression in patients with AS, without increasing toxicity importantly.

# Introduction

In patients with Ankylosing Spondylitis (AS) numerous studies have demonstrated that non steroidal anti-inflammatory drugs (NSAIDs) provide rapid relief of inflammatory back pain and stiffness, and improve physical function[1-6]. NSAIDs belong to the most frequently prescribed drugs in AS, but gastrointestinal toxicity limits their long-term use. Gastrointestinal adverse events are associated with the inhibition of cyclo-oxygenase-1 (COX-1), which is responsible for the production of cytoprotective prostaglandins in the gastric mucosa[7]. The latest generation of NSAIDs selectively inhibits COX-2, that is upregulated under inflammatory conditions and responsible for the production of pro-inflammatory prostaglandins, and leaves COX-1 relatively undisturbed. COX-2 selective NSAIDs are associated with a reduced risk of serious gastrointestinal complications[8,9], but are at least similarly effective in AS as conventional NSAIDs[2,3].

Because of the risk of serious adverse events, many physicians currently recommend their patients with AS to take NSAIDs only if necessary. It is, however, not established whether the use of NSAIDs can alter the long-term outcome of the disease. Most studies to the efficacy of NSAIDs have been short-term studies (up to 6 weeks duration). Measures of spinal mobility and levels of acute-phase reactants generally did not improve in these studies[3,6,10,11]. One placebo controlled 1-year study compared piroxicam with low – and high dose meloxicam, and with placebo[1]. At 1 year, and not at 6 weeks, significantly more improvement was observed in chest expansion and C-reactive protein as compared to placebo. These findings suggest that NSAIDs may to some extent control the disease process. Another study that points to the disease controlling potential of NSAIDs is an uncontrolled retrospective observational study, showing that phenylbutazone impaired the ossification of the vertebral column in patients with AS[12].

The improved gastroprotective safety profile of COX-2 selective- as compared to unselective - NSAIDs justifies a formal test of the hypothesis that NSAIDs may alter the course of AS, thus disputing the current recommendation to take NSAIDs only on demand. In a 2-year randomized clinical trial we compared the strategies long-term continuous NSAID-use, and NSAID-use on demand only, with respect to their influence on radiographic progression.

# Methods

## Patients

The study was conducted from June 1998 to July 2001 by rheumatologists from 75 centers, both hospitals and private practices in France. The Ethics Committee of Cochin Hospital, Paris, approved the study. Two hundred and fifteen outpatients fulfilling the modified New York criteria for AS[13], who had previously participated in a 6-week

randomized, double-blind clinical trial, which compared celecoxib 100 mg twice daily to ketoprofen 100 mg twice daily, and to placebo, were included after written informed consent[3]. Criteria for inclusion in the 6-week trial were: 1) A daily NSAID intake during the month preceding the screening visit; 2) a NSAID washout period of 2-14 days before the baseline visit; and 3) a flare of the disease at baseline, defined as a) an absolute pain score ≥40 mm on a 100-mm visual analogue scale (VAS); and b) increase in pain of at least 30%, between the screening and baseline visits. Patients with peripheral arthritis, defined as the presence of active (with swelling) synovitis of a peripheral joint (excluding the shoulder) at the screening visit, and those with active inflammatory bowel disease, were excluded, as well as those with severe concomitant medical illnesses. Patients who had received corticosteroids during the previous 6 weeks, and / or any disease-modifying antirheumatic drug with a dose-change during the previous 6 months, were also excluded, as well as patients with peptic ulcer disease confirmed by gastro-duodenoscopy within the year preceding the screening visit.

### Study design

The present study was a randomized, open label, comparative trial of two strategies. At the final visit of the preceding 6-week trial, which is considered the baseline visit of the present study, patients were randomly allocated to a continuous treatment strategy, or to a treatment strategy with drugs on demand only. Randomization was performed by a computer-generated randomization list. Patients allocated to the continuous treatment strategy were treated with a NSAID irrespective of symptoms, for a period of two years. Patients allocated to the on demand treatment strategy were instructed to take their NSAID only in case of serious symptoms (pain, stiffness). Patients in both treatment groups started with celecoxib 100mg twice daily, and were allowed to increase the dose to 200 mg twice daily if necessary. It was at the discretion of the patient to decide to increase the dose. If for any reason (inefficacy or adverse event) the initial treatment with celecoxib was discontinued, patients were allowed to continue the study with any other NSAID, but they were instructed to maintain the allocated treatment strategy (continuous or on demand only). Compliance was assessed by pill count.

### Study visits

There were 10 planned visits (figure 7.1), as follows: baseline visit (M0), one follow-up visit after an interval of one month (M1), 7 follow-up visits (M4, M7, M10, M13, M16, M19, M22) with an interval of 3 months each, and a closing visit at 24 months (M24). At every visit clinical signs and symptoms and adverse events were assessed. Laboratory tests were performed at visit M1, and at the visits M7, M13, M19 and M24. Radiography of the spine was performed at baseline and at 24 months.

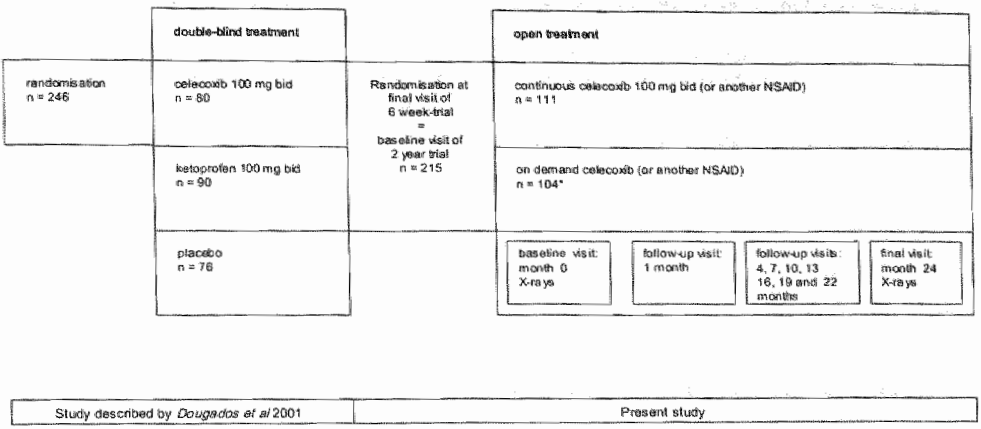



Figure 7.1 Flow chart

## Assessments

Structural damage was scored on radiographs of the lumbar- and cervical spine according to the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS)[14] by one observer (AW) who was blind for the treatment strategy and for the time order of the radiographs. The difference between the mSASSS at M0 and M24 was considered the progression score (range 0 to 72). Intra- and inter-observer reliability of the mSASSS were tested in a previous experiment. Intra-class correlation coefficients were 0.95 and 0.82 respectively[14].

Disease activity was measured by the 6-question patient-reported Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) (0: no disease activity, to 100: highest level of disease activity)[15]. Functional capacity was measured by the 10-question patient-reported Bath Ankylosing Spondylitis Functional Index (BASFI) (from 0: lowest level, to 100: highest level)[16]. Pain was measured by three variables, as follows: 1) the global intensity of pain (visual analogue scale (VAS) from 0: no pain, to 100: extreme pain); 2) the spinal pain index consisting of 4 items: pain in the cervical spine, pain in the dorsal spine, pain in the lumbar spine and pain in the SI joints (each item was assessed on a Likert scale ranging from 0: no pain, to 4: unbearable pain). The spinal pain index was calculated as the sum score of the 4 items (from 0: no pain, to 16: extreme pain); and 3) the percentage of days with pain during the last three months (VAS from 0: no pain at all during 3 months, to 100: pain every day during 3 months). Inflammation was measured by 3 variables, as follows: 1) nocturnal pain during the last week (VAS from 0: no nocturnal pain, to 100: extreme nocturnal pain; 2) duration and severity of morning

stiffness during the last week (score that represents the mean of the 5th and 6th BASDAI questions (VAS)); and 3) the level of C-reactive protein (CRP) and the erythrocyte sedimentation rate (ESR). Spinal mobility was measured by 4 variables, as follows: 1) fingertip to floor distance (in cm); 2) modified Schober (in cm); 3) chest expansion (in cm); and 4) occiput to wall distance (in cm)[17]. Fatigue was measured by the first question of the BASDAI (VAS from 0: no fatigue, to 100: extreme fatigue). Global disease activity during the last week was measured both by the patient and the investigator (VAS from 0: no disease activity, to 100: severe disease activity).

### Missing variable values

Missing values of variables assessing signs and symptoms were replaced by the last observation that was present, which was carried forward, provided that at least one value obtained under treatment was available.

### Safety variables

The investigator actively asked for adverse events at every visit. Laboratory assessments including haemoglobin, platelet count, serum creatinine, ASAT and ALAT were performed at 6-month intervals. Systolic and diastolic blood pressure were measured every visit.

### Analysis and statistics

An *exploratory analysis* included time-plots of all variables assessing signs and symptoms, stratified by treatment group, and probability plots of radiographic progression scores, stratified by treatment group. Probability plots present every patient's progression score against its cumulative frequency (expressed as a proportion; cumulative probability[18]).

The primary set of s*tatistical analyses* included a between group comparison of: 1) radiographic progression scores (by Mann-Whitney U-test); 2) time-averaged values for variables reflecting signs and symptoms (by linear regression analysis with treatment group and baseline values of these variables as independent variables); and 3) the frequency of reported site-specific adverse events (by Chi-square test, or Fisher's exact test if appropriate).

Other statistical analyses included: 1) linear regression analysis on radiographic progression, with treatment group as factor and baseline variables of disease activity and radiographic damage as covariate (to investigate whether baseline difference could account for between-group differences at the end of the trial); 2) linear regression analysis on radiographic progression with treatment group as factor, and time averaged means of disease activity variables as covariates (to investigate whether a difference in radiographic progression could be explained by differences in disease activity during follow-up). Van der Waerden-normalized data were used in the linear regression analyses if variables had a non-normal distribution pattern. Residual plots were checked for homogeneity and linearity.

A sensitivity analysis included reanalysis of the primary results after imputation of missing data by two means: 1) imputation by the mean value of the total population (which overestimates true progression); and 2) imputation by 0 (which underestimates true progression).

This study was performed with an unrestricted grant from Pharmacia, France.

# Results

## Patients

Two hundred and fifteen patients were randomized, 111 to the continuous treatment group and 104 to the on demand treatment group. One patient in the on demand group was excluded from the analysis because he died in a car-accident before starting the trial.

In the continuous treatment group, 96 patients completed the study: 68 patients completed while using celecoxib, and 28 patients while using a different NSAID. The reasons for withdrawal of the 15 patients in this group were; inefficacy in eight patients, adverse events in two patients, moving to another city or country in two patients, and the reason was unknown in three patients.

In the on demand treatment group, 86 patients completed the study: 67 patients completed while using celecoxib, and 19 patients while using a different NSAID. The reasons for withdrawal of the 17 patients in this group were; inefficacy in eight patients, adverse events in three patients, moving to another city or country in two patients, and the reason was unknown in four patients.

Complete sets of radiographs were available in 76 patients in the continuous group and 74 patients in the on demand group. The baseline characteristics of patients in both treatment groups are shown both for all randomized patients, as well as for the patients with a complete set of radiographs available (table 7.1). With respect to all randomized patients, between-group differences at baseline were small and negligible. With respect to patients with a complete set of radiographs available, age and disease duration were lower in patients in the on demand group, as compared to patients in the continuous group, but these differences were not statistically significant. Other baseline variables were similar.

The values of clinical measurements at baseline are shown for all randomized patients (table 7.2), and for those with complete radiographs available separately (table 7.3). Overall, disease activity tended to be lower in the continuous group as compared to the on demand group, and this difference was more pronounced in the selection of patients with complete sets of radiographs available, but the between-group differences were not statistically significant.

Table 7.1 Baseline characteristics by treatment group

| Characteristic | All patients | | Patients with a complete set of X-rays | |
|---|---|---|---|---|
| | continuous use (n=111) | on demand use (n=103) | continuous use (n=76) | on demand use (n=74) |
| Age (years) | 38.0 (10.7)* | 40.1 (10.5) | 40.9 (9.8) | 37.9 (11.9) |
| Male (%) | 67 | 72 | 66 | 70 |
| Disease duration (years) | 11.9 (9.3) | 11.0 (9.4) | 13.0 (10.2) | 10.2 (9.3) |
| HLA-B27 positive (%) | 86 | 87 | 88 | 88 |
| DMARD use (%) | 29 | 26 | 26 | 27 |
| Sulfasalazine use | 25 | 22 | 24 | 27 |
| MTX use | 3 | 2 | 1 | 3 |
| Other** | 2 | 2 | 1 | 3 |
| Analgetics use (%) | 10 | 9 | 11 | 9 |

* mean (SD), **Gold or corticosteroids

Table 7.2 Disease activity at baseline, and during the trial, for all patients.

| | Baseline | | Month 1 | | Time averaged means | | |
|---|---|---|---|---|---|---|---|
| | continuous (n=111) | on demand (n=103) | continuous (n=111) | on demand (n=103) | continuous (n=111) | on demand (n=103) | p |
| **Disease activity** | | | | | | | |
| BASDAI (0–100) | NA | NA | 30 (19) | 32 (24) | 30 (18) | 32 (20) | 0.51 |
| **Function** | | | | | | | |
| BASFI (0–100) | 33 (25) | 38 (28) | 31 (24) | 32 (26) | 30 (21) | 31 (23) | 0.33 |
| **Pain** | | | | | | | |
| Global pain (0–100) | 50 (38) | 54 (37) | 37 (27) | 39 (27) | 37 (22) | 40 (23) | 0.44 |
| Spinal pain index (0–16) | 6.4 (3.6) | 6.9 ( 3.3) | 5.5 ( 3.6) | 6.1 (3.3) | 5.4 (3.2) | 5.7 (2.8) | 0.88 |
| Percentage painful days (0–100) | NA | NA | 46 (33) | 52 (35) | 45 (27) | 49 (26) | 0.32 |
| **Inflammation** | | | | | | | |
| Night pain (0 – 100) | 38 (32) | 43 (34) | 25 (25) | 31 (28) | 27 (21) | 32 (24) | 0.91 |
| BASDAI morning *stiffness (0–100)* | NA | NA | 30 (22) | 34 (27) | 29 (20) | 32 (22) | 0.61 |
| CRP (mg/l) | 14.7 (17.9) | 12.7 (17.1) | 15.8 (22.6) | 12.0 (15.5) | 14.5(17.8) | 12.3 (16.1) | 0.82 |
| ESR (mm/hour) | 17.0 (13.8) | 17.0 (16.7) | 16.3 (13.7) | 16.9 (15.7) | 15.7 (11.4) | 15.8 (13.3) | 0.40 |
| **Spinal mobility** | | | | | | | |
| Finger to floor distance (cm) | 21.2 (15.6) | 23.0 (14.7) | 19.4 (14.9) | 19.4 (14.1) | 19.0 (13.6) | 19.7 (13.2) | 0.48 |
| Schober (cm) | 3.2 (1.4) | 3.2 (1.4) | 3.2 (1.5) | 3.3 (1.4) | 3.2 (1.3) | 3.3 (1.3) | 0.97 |
| Chest expansion (cm) | 4.7 (2.3) | 5.0 (2.3) | 4.8 (2.2) | 5.2 (2.1) | 5.0 (2.2) | 5.3 (2.1) | 0.65 |
| Occiput to wall (cm) | NA | NA | 3.2 (4.3) | 3.0 (3.6) | 3.5 (4.5) | 2.8 (3.3) | 0.51 |
| **Fatigue (0–100)** | NA | NA | 38 (24) | 38 (28) | 38 (22) | 40 (24) | 0.53 |
| **Global Assessment** | | | | | | | |
| Patient global assessment (0–100) | 43 (29) | 47 (31) | 37 (27) | 40 (26) | 37 (23) | 40 (22) | 0.94 |
| Physician global assessment (0–100) | 42 (28) | 44 (29) | 32 (25) | 35 (25) | 32 (20) | 34 (19) | 0.60 |

Values are the mean (standard deviation) P: between group differences were analysed by linear regression with treatment group and baseline values (if not available M1 scores) as independent variables. BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; BASFI = Bath Ankylosing Spondylitis Functional Index; CRP = C-reactive protein; ESR = Erythrocyte Sedimentation Rate; NA = not available.

Table 7.3 Disease activity at baseline, and during the trial, for patients with a complete set of X-rays available.

| | Baseline | | Month 1 | | Time averaged means | | |
|---|---|---|---|---|---|---|---|
| | continuous (n=76) | on demand (n=74) | continuous (n=76) | on demand (n=74) | continuous (n=76) | on demand (n=74) | p |
| **Disease activity** | | | | | | | |
| BASDAI (0-100) | NA | NA | 29 (18) | 31 (25) | 26 (17) | 32 (20) | 0.17 |
| **Function** | | | | | | | |
| BASFI (0-100) | 30 (23) | 36 (28) | 28 (22) | 31 (26) | 27 (19) | 30 (23) | 0.72 |
| **Pain** | | | | | | | |
| Global pain (0-100) | 46 (37) | 52 (37) | 35 (25) | 37 (26) | 33 (19) | 39 (22) | 0.13 |
| Spinal pain index (0-16) | 5.7 (3.3) | 6.6 (3.0) | 5.0 (3.3) | 5.9 (3.3) | 4.6 (2.7) | 5.8 (2.8) | 0.07 |
| Percentage painful days (0-100) | NA | NA | 45 (33) | 54 (35) | 39 (23) | 48 (25) | 0.12 |
| **Inflammation** | | | | | | | |
| Night pain (0-100) | 34 (31) | 41 (33) | 21 (22) | 32 (29) | 22 (18) | 32 (24) | 0.01 |
| BASDAI morning stiffness (0-100) | NA | NA | 29 (21) | 33 (27) | 26 (19) | 31 (21) | 0.25 |
| CRP (mg / dl) | 13.1 (15.3) | 12.2 (17.5) | 13.7 (22.8) | 10.8 (13.4) | 12.8 (14.8) | 16.5 (14.4) | 0.82 |
| ESR (mm/ first hour) | 16.5 (13.1) | 17.5 (18.1) | 16.3 (13.6) | 17.1 (16.7) | 15.0 (10.4) | 12.2 (15.8) | 0.40 |
| **Spinal mobility** | | | | | | | |
| Finger to floor distance (cm) | 18.0 (14.4) | 21.3 (13.2) | 17.7 (14.7) | 17.9 (13.3) | 16.1 (13.0) | 18.3 (12.3) | 0.76 |
| Schober (cm) | 3.3 (1.4) | 3.2 (1.4) | 3.3 (1.4) | 3.2 (1.4) | 3.4 (1.3) | 3.2 (1.3) | 0.31 |
| Chest expansion (cm) | 4.9 (2.3) | 5.2 (2.3) | 5.0 (2.2) | 5.3 (2.1) | 5.3 (2.2) | 5.5 (2.0) | 0.68 |
| Occiput to wall (cm) | NA | NA | 2.8 (3.7) | 2.7 (3.7) | 2.9 (4.2) | 2.5 (3.3) | 0.45 |
| **Fatigue (0-100)** | NA | NA | 36 (25) ) | 36 (28) | 33 (19) | 38 (23) | 0.02 |
| **Global assessment** | | | | | | | |
| Patient global assessment (0-100) | 39 (27) | 44 (31) | 34 (24) | 39 (27) | 31 (20) | 39 (22) | 0.07 |
| Physician global assessment (0-100) | 38 (26) | 41 (29) | 30 (24) | 34 (25) | 27 (17) | 34 (19) | 0.05 |
| *mSASSS score* | 7.9 (14.7) | 9.3 (15.2) | | | | | |

Values are the mean (standard deviation). P: between group differences were analysed by linear regression with treatment group and baseline values if not available M1 scores) as independent variables. BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; BASFI = Bath Ankylosing Spondylitis Functional Index; CRP = C-reactive protein; ESR = Erythrocyte Sedimentation Rate.;NA = not available.

## Mean daily dose of celecoxib

Based on the pill counts, the mean dose of celecoxib taken per day was 243 mg (SD 59) in the continuous group and 201 mg (SD 93) in the on demand group. The difference of 42 mg 95% CI (21–63) was statistically significant (p=0.0001).

### Radiographic progression

The mean (SD) baseline radiographic damage score in both groups was similar, 7.9 (14.7) mSASSS units in the continuous treatment group and 9.3 (15.2) units in the on demand treatment group.

The probability plot of the mSASSS progression scores for the two treatment groups over 24 months (figure 7.2) showed radiographic progression in a greater proportion of patients in the on demand treatment group (45%), as compared to the continuous treatment group (22%). Using a cut-off of ≥3 units, again twice as many patients in the on demand group compared to the continuous group show this level of progression (23% vs. 11%). The maximum progression scores are 4 and 9 units in the continuous group and on demand group respectively. The curve of the on demand group lies left to the curve of the continuous group along the entire range, reflecting a higher level of radiographic progression. The mean (SD) radiological progression after two years was 0.4 (1.7) mSASSS-units in the continuous treatment group, and 1.5 (1.7) mSASSS-units in the on demand group. The between-group difference was statistically significant (p=0.002).

Sensitivity analysis showed that imputation of missing data by different means did not influence the direction of the between-group difference (p=0.002 for the between-group difference after imputation with the entire group-mean; and p=0.077 after imputation with the value 0).

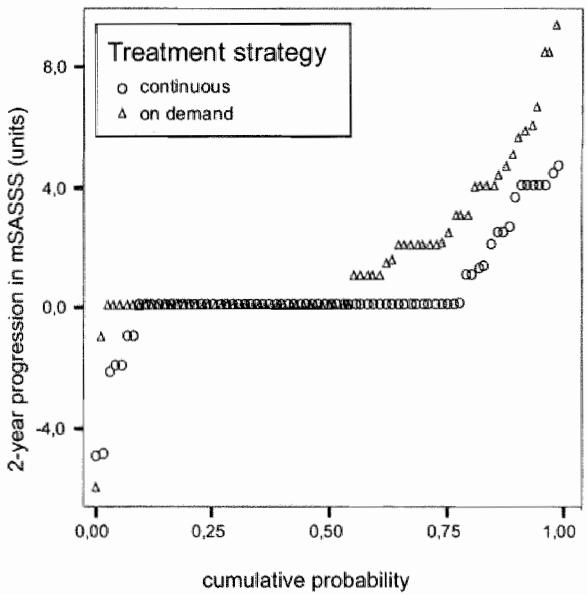



Figure 7.2 Probability plot of mSASSS progression scores over 24 months

### Signs and symptoms

The course of four variables of disease activity, which can be considered representative for the other variables including those reflecting spinal mobility, is shown in figure 7.3. It is obvious that disease activity is stable over time in both groups, after some decrease in global pain between baseline and the first assessment. The general impression is that disease activity, measured by any of these variables, is somewhat higher in the on demand group, as compared to the continuous group. Time-averaged values of all variables reflecting signs and symptoms are shown in tables 7.2 and 7.3. The results confirm the impression of a somewhat higher disease activity in the on demand as compared to the continuous group, but the differences were not statistically significant for any variable. When limited to the patients with a complete set of radiographs (table 7.3), the time-averaged values of night pain (p=0.01), fatigue (p=0.02) and physician's global assessment (p=0.05) were significantly worse in the on demand treatment group. Adjustments for multiple testing were not performed.

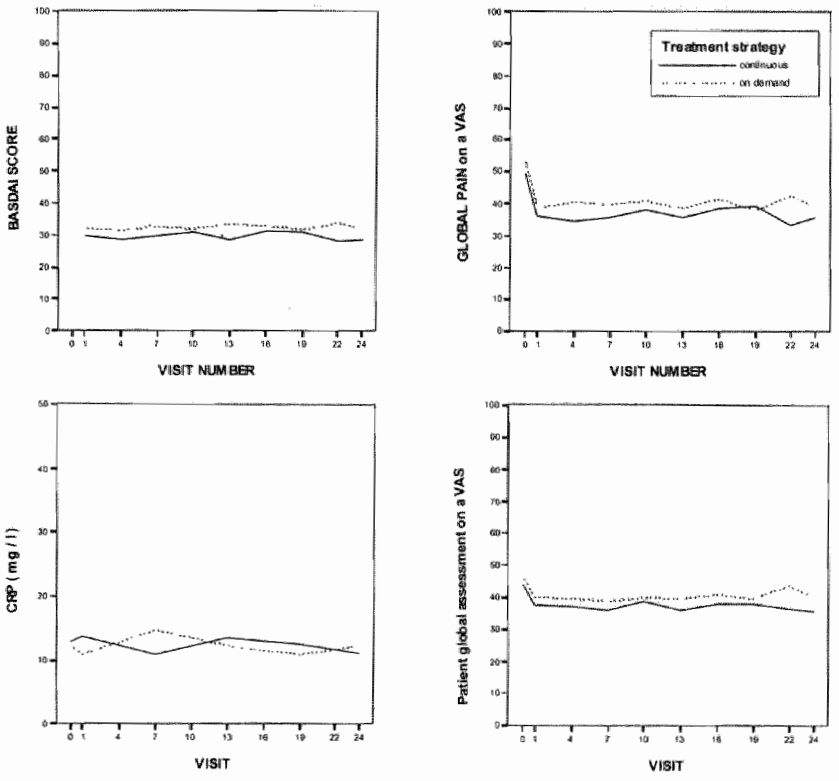



Figure 7.3 Status scores of BASDAI, global pain score, CRP, and patient global assessment for the patients with a complete set of X-rays by treatment strategy (means, SE). BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; VAS = Visual Analogue Scale; CRP = C-reactive protein.

### Confounding

We investigated in the group of patients with complete radiographs whether the observed baseline differences in signs and symptoms could explain the between-group difference in radiographic progression, by performing linear regression analysis on radiographic progression, with treatment as factor and baseline values of variables reflecting signs and symptoms (all entered separately) as covariates. The factor treatment remained statistically significant in all analyses, and the regression coefficient for treatment did not change importantly (always <10%), which indicates that the between-group difference in radiographic progression cannot be explained by baseline differences.

We further investigated whether differences in signs and symptoms *during* treatment could explain the between-group difference in radiographic progression, by entering the time-averaged values as covariate in the linear regression analysis mentioned before.

Again, the regression coefficient for the factor treatment was not importantly influenced by any of the time averaged variables, suggesting that differences in signs and symptoms during follow up could not explain the observed difference in radiographic progression.

### Safety

Serious adverse events (SAEs)were reported 22 times in 22 separate patients (29.8%) of the continuous group, and 25 times in 16 separate patients (19.8%) of the on demand group. Only one of these SAEs (a case of severe abdominal pain requiring hospital admission in the on demand group) was considered related to the study medication by the treating physician.

The most important and frequent adverse events are reported in table 7.4.

Gastrointestinal adverse events occurred most frequently, followed by respiratory tract infections and skin rashes. Numerically, some relevant adverse events, such as hypertension (in 10 vs. 3 patients; p=0.12), abdominal pain (in 12 vs. 6 patients; p=0.28), diarrhea (in 21 vs. 13 patients; p=0.28) and dyspeptic complaints (in 46 vs. 39 patients; p=0.65) occurred more frequently in the continuous group as compared to the on demand group, but the differences were not statistically significant. Depressive symptoms occurred more frequently in the continuous group (in 15 vs. 4 patients; p=0.03), and this difference was statistically significant.

The mean blood pressure remained at a constant level, and was similar in both groups. Hemoglobin, ASAT, ALAT and serum creatinine all remained at a constant level, and were similar in both groups.

Table 7.4. Number of adverse events per treatment group

| Type of adverse event | Adverse event | Treatment group continuous use (n=111) | on demand use (n=103) |
|---|---|---|---|
| Cardiovascular | Hypertension | 10 | 3 |
| | Angina pectoris | 0 | 2 |
| | Coronary artery disorder | 0 | 1 |
| | Myocardial infarction | 0 | 2 |
| | Edema | 4 | 3 |
| Neurological | Headache | 13 | 13 |
| | Vertigo | 7 | 5 |
| Gastro-intestinal | Abdominal pain | 12 | 6 |
| | Diarrhoea | 21 | 13 |
| | Duodenal ulcer | 1 | 0 |
| | Dyspeptic complaints | 46 | 39 |
| | Gastroenteritis | 17 | 13 |
| | Esophageal complaints | 5 | 4 |
| Laboratory abnormalities | Increased liver enzymes | 3 | 0 |
| | Increased serum creatinine level | 1 | 0 |
| Neoplastic | Undifferentiated carcinoma | 0 | 2 |
| | Gastro-intestinal malignant neoplasm | 0 | 1 |
| | Pseudo mononucleosis | 1 | 0 |
| Platelet, bleeding & clotting disorders | | 1 | 3 |
| Psychiatric | Anxiety | 7 | 6 |
| | Depressive symptoms | 15 | 4 |
| | Sleeping disorder | 8 | 5 |
| Respiratory | Upper respiratory tract infections | 45 | 44 |
| | Lower respiratory tract infections | 14 | 17 |
| Skin | Rash | 5 | 8 |
| | Pruritis | 9 | 6 |

# Discussion

The main conclusion of this trial, which compared the strategies continuous NSAID-use and NSAID-use on demand only, is that a continuous treatment strategy reduces radiographic progression despite a similar effect of both strategies on signs and symptoms (pain, inflammation, spinal mobility), whereas a strategy of continuous NSAID-use is not associated with significantly more toxicity. This observation provides a strong indication that NSAIDs may have disease controlling properties.

The ASsesment in Ankylosing Spondylitis (ASAS) working group selected radiographic evaluation as an obligatory outcome assessment to prove disease controlling properties[19]. But until recently, and in contrast with the situation in rheumatoid arthritis (RA), only very few studies had included radiology as an outcome parameter. A literature search for the potential of NSAIDs to reduce radiographic progression revealed only one article: Boersma[12] who performed an uncontrolled cohort study

suggested already in 1975 that continuous use of phenylbutazone may reduce the ossification of the vertebral column in patients with AS. The best explanation for this scarceness of data is that NSAIDs were considered symptom modifiers rather than drugs that may control the course of the disease. The picture changed after the demonstration that TNF-blocking drugs, that inhibit radiographic progression in RA, are also highly effective in alleviating the symptoms of AS[20,21], and it is expected that in analogy to RA TNF-blocking drugs can inhibit radiographic progression in AS. Another reason to explain the absence of radiographic data in AS clinical trials is the lack of a reliable scoring method. Only recently, we have demonstrated the superiority (with respect to sensitivity to change) of the mSASSS in detecting 2-year change in a method-comparing cohort study[14], and we decided to use this scoring method in the present study. Indeed, it appeared to be possible to detect progression after 2 years in a significant proportion of patients in both groups by using the mSASSS, scored by one single reader who scored with concealed time order. We feel concealment of reading order is as important as scoring blind for treatment allocation, because it prevents expectation bias, and the occurrence of negative scores gives a good impression about the level of measurement error (under the assumption that true negative scores are impossible)[18]. As such, concealment of time order and treatment allocation both add to the validity of the results, and to their credibility.

The finding that NSAIDs reduce radiographic progression requires a proper biological explanation. COX-2 is relevant in bone formation: both COX-2 knock-out mice and mice treated with COX-2 inhibiting drugs showed reduced callus formation after a fracture, which is due to suppression of osteoblasts[22]. In an immunohistochemical study, comparing synovial samples of patients with osteoarthritis, RA, psoriatic arthritis and AS, COX-2 expression appeared to be highest in AS synovial samples[23]. If an up-regulated level of COX-2 in AS is indeed responsible for increased osteoblastic bone formation (syndesmophytes), inhibition of COX-2 by NSAIDs may be a rational approach to prevent the occurrence of syndesmophytes. Both non-selective- and COX-2 selective NSAIDs inhibit COX-2, and radiographic effects of COX-2 inhibition in AS can therefore be expected with selective and non-selective NSAIDs. The clinical finding that (conventional) NSAIDs may reduce the risk of heterotopic bone formation after hip-arthroplasty by 50-65%[24,25] is in line with the observations in animal models, as well as with the results of our study.

It could be argued that if a tighter control of disease activity in the continuous group was strived for, the difference might have been even greater. However, we noticed a remarkable lack of association between radiographic progression and variables reflecting disease activity (pain and inflammation), which is in contrast with the situation in RA[26,27]. Data of our study suggest – but do not prove - that inflammation and progression of structural damage are two separate processes in AS. A study by Lussier et al in a rat experimental model provides support for such a dissociation[26]: of 3 different

NSAIDs, the drug exhibiting almost no anti-inflammatory activity (phenylbutazone) appeared to be the best inhibitor of ossification.

We found that the incidence of a number of relevant adverse events in the continuous treatment group was higher (although not statistically significant) than in the on demand treatment group. It is important to mention, however, that this study was not designed to get a reliable picture of adverse events associated with long-term NSAID-use: patients were not blinded for treatment strategy, and knowledge of the treatment strategy may well have influenced the report rate of adverse events. Given the rather low incidence of relevant adverse events in the context of the design of the study, and given the absence of drug-related serious adverse events, we conclude that both strategies have an acceptable long-term toxicity profile.

This study may evoke a number of concerns. First, a trial comparing strategies, in which drugs and doses are not completely fixed, is susceptible to bias because patients and physicians try to minimize the level of pain within the limits of the protocol (confounding by indication). But this type of confounding will generally reduce between-group differences, and cannot be responsible for the contrast in radiographic progression. Second, only the dose of celecoxib was recorded, and not the dose of other NSAIDs (after a switch). We were therefore not able to check whether patients were still compliant to the allocated treatment strategies if they had switched NSAIDs. But again, a bias caused by violation of the allocated treatment strategy would have obscured rather than revealed a true between-group contrast in radiographic progression. The between-group difference in mean dose of celecoxib was rather small (although highly statistically significant). It should be noted, however, that a mean dose does not appropriately reflect the pattern of drug use: strategies of moderate doses of a NSAID used daily, and high doses of a NSAID used every other day, may both arrive at the same mean dose, but may have different pharmacodynamic consequences. As such, the mean celecoxib dose is not the best parameter to explain the differences in radiographic progression. Additional information on the timing of dosing would have shed more light on this aspect. However, this information cannot be deducted from pill counts and the only valid instrument to measure this is by an electronic monitoring device which registers date and time of opening the pill bottle. However, this technique was not applied in the present study.
Third, one may think that the observation of reduced progression is coincidental (Type I error), or biased by baseline differences, and/or missing observations in 30% of the patients. Obviously, Type I error can never be entirely excluded, and our observation awaits confirmation from other studies. But several arguments point to a true effect: the sensitivity analyses by various means of missing values imputation all arrived at the same result; the treatment contrast did not disappear after adjustment for any of the set of potential confounders at baseline; at last, there are arguments for the biological plausibility of our observation in the literature, which may add to its validity.

Thus, we here conclude that a strategy of continuous use of NSAIDs reduces radiographic progression in patients with AS, without importantly increasing toxicity. Awaiting confirmation of these results, we carefully recommend that, if patients need NSAIDs to reduce signs and symptoms of AS, they use it continuously instead of on demand. Currently, data underscoring such a recommendation in asymptomatic patients is lacking.

# References


1. Dougados M, Gueguen A, Nakache JP, Velicitat P, Veys EM, Zeidler H, Calin A. Ankylosing spondylitis: what is the optimum duration of a clinical study? A one year versus a 6 weeks non-steroidal anti-inflammatory drug trial. Rheumatology (Oxford) 1999;38:235-44.
2. Melian A, van der Heijde DMFM, James MK, Calin A, Giallella K, Reicin A, et al. Etoricoxib in the treatment of Ankylosing Spondylitis. ACR annual scientific meeting 2002 abstract 1131.
3. Dougados M, Behier JM, Jolchine I, Calin A, van der Heijde D, Olivieri I, Zeidler H, Herman H. Efficacy of celecoxib, a cyclooxygenase 2-specific inhibitor, in the treatment of ankylosing spondylitis: a six-week controlled study with comparison against placebo and against a conventional nonsteroidal antiinflammatory drug. Arthritis Rheum 2001;44:180-5.
4. Dougados M. Treatment of spondyloarthropathies. Recent advances and prospects in 2001. Joint Bone Spine 2001;68:557-63.
5. Calin A, Grahame R. Double-bind cross-over tiral of flurbiprofen and phenylbutazone in ankylosing spondylitis. Br Med J 1974;4:496-9.
6. Dougados M, Nguyen M, Caporal R, Legeais J, Bouxin-Sauzet A, Pellegri-Guegnault B, Gomeni C. Ximoprofen in ankylosing spondylitis. A double blind placebo controlled dose ranging study. Scand J Rheumatol 1994;23:243-8.
7. Dougados M, Dijkmans B, Khan M, Maksymowych W, van der Linden S, Brandt J. Conventional treatments for ankylosing spondylitis. Ann Rheum Dis 2002;61 Suppl 3:iii40-50.
8. Langman MJ, Jensen DM, Watson DJ, Harper SE, Zhao PL, Quan H, Bolognese JA, Simon TJ. Adverse upper gastrointestinal effects of rofecoxib compared with NSAIDs. JAMA 1999;282:1929-33.
9. Simon LS, Weaver AL, Graham DY, Kivitz AJ, Lipsky PE, Hubbard RC, Isakson PC, Verburg KM, Yu SS, Zhao WW, Geis GS. Anti-inflammatory and upper gastrointestinal effects of celecoxib in rheumatoid arthritis: a randomized controlled trial. JAMA 1999;282:1921-8.
10. Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988;15:302-7.
11. Dougados M, Caporal R, Doury P, Thiesce A, Pattin S, Laffez B, Amor B. A double blind crossover placebo controlled trial of ximoprofen in as. J Rheumatol 1989;16:1167-9.
12. Boersma JW. Retardation of ossification of the lumbar vertebral column in ankylosing spondylitis by means of phenylbutazone. Scand J Rheumatol 1976;5:60-4.
13. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27: 361-8.
14. Wanders AJB, Landewé RBM, Spoorenberg A, Dougados M, van der Linden, S Mielants H, van der Tempel H, van der Heijde DM. What is the most appropriate radiologic scoring method in Ankylosing spondylitis clinical trials. A comparison of the available methods based on the OMERACT filter. Arthritis Rheum 2004;50:2622-32.
15. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994;21:2286-91.
16. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, Jenkinson T. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281-5.
17. Bellamy N. Clinimetric concepts in outcome assessment: the OMERACT filter. J Rheumatol 1999;26:948-50.
18. Landewé R, Van der Heijde DFMF. Radiographic progression visualised depicted by probability plots: presenting data with optimal use of the individual values. Arthritis Rheum.2004; 50:699-706.

19. van der Heijde D, van der Linden S, Bellamy N, Calin A, Dougados M, Khan MA. Which domains should be included in a core set for endpoints in ankylosing spondylitis? Introduction to the ankylosing spondylitis module of OMERACT IV. J Rheumatol 1999;26:945-7.
20. Brandt J, Haibel H, Cornely D, Golder W, Gonzalez J, Reddig J, Thriene W, Sieper J, Braun J. Successful treatment of active ankylosing spondylitis with the anti-tumor necrosis factor alpha monoclonal antibody infliximab. Arthritis Rheum 2000;43:1346-52.
21. Gorman JD, Sack KE, Davis JC, Jr. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor alpha. N Engl J Med 2002;346:1349-56.
22. Zhang X,Schwarz EM, Young DA, Puzas JE, Rosier RN, O'Keefe RJ. Cyclooxygenase-2 regulates mesenchymal cell differentation into the osteoblast lineage and is critically involved in bone repair. J Clin Invest 2002;109:1405-15.
23. Siegle I, Klein T, Backman JT, Saal JG, Nusing RM, Fritz P. Expression of cyclooxygenase 1 and cyclooxygenase 2 in human synovial tissue: differential elevation of cyclooxygenase 2 in inflammatory joint diseases. Arthritis Rheum 1998 41: 122-9.
24. Ijiri K, Matsunaga S, Fukuda T, Shimizu T. Indomethacin inhibition of ossification induced by direct current stimulation. J Orthop Res 1995;13:123-31.
25. Neal B, Rodgers A, Dunn L, Fransen M. Non-steroidal anti-inflammatory drugs for preventing heterotopic bone formation after hip arthroplasty. Cochrane Database Syst Rev 2000: CD001160.
26. Wolfe F. The prognosis of rheumatoid arthritis: assessment of disease activity and disease severity in the clinic. Am J Med 1997;103:12S-18S.
27. Lindqvist E, Jonsson K, Saxne T, Eberhardt K. Course of radiographic damage over 10 years in a cohort with early rheumatoid arthritis. Ann Rheum Dis 2003; 62(7):611-6.
28. Lussier A, de Medicis R. Correlation between ossification and inflammation using a rat experimental model. J Rheumatol Suppl 1983;11:114-7.

# Chapter 8

## Summary in perspective

# Summary in perspective

The studies described in this thesis cover important aspects of outcome assessment in Ankylosing Spondylitis (AS). For decades, research in AS was focused on eliciting the pathophysiological mechanisms operative in this heterogeneous disease, rather than on outcome assessment. The latter was considered unnecessary, since effective drugs were lacking and clinical trials were rarely performed in AS.
Two consecutive developments have changed this picture dramatically. The first development was the foundation of the international ASsessment in Ankylosing Spondylitis (ASAS) working group in 1995 in Amsterdam, the subsequent expert-based establishment of domains of outcome, and core sets of measurements to assess efficacy within these domains, and the endorsement of these core sets by the Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) initiative in 1998. The second development was the discovery of biological drugs that have the potential to inhibit (the activity of) tumor necrosis factor (TNF)-alpha, and their introduction in chronic inflammatory diseases such as RA. Since anecdotal reports suggested efficacy of TNF-blocking drugs in AS and in Chrohn's disease (a disease associated with AS) as well, and pharmaceutical industry expressed their willingness to further develop TNF-blocking drugs in this disease, an immediate need for appropriate outcome measures to be used in clinical trials arose. The application of TNF-blocking drugs in AS clinical trials has boosted the recent development of outcome measures in AS, has been a justification for the scientific work summarized in this thesis, and has set the research agenda for the forthcoming years.

In chapter 2, responsiveness and discriminatory capacity of measurement instruments included in the core set for disease controlling antirheumatic therapy (DC-ART) in AS was investigated, making use of the data of one of the first clinical trials with TNF-blocking therapy in AS. Responsiveness and discriminatory capacity are important characteristics of measurement instruments with respect to their use in clinical trials. It was shown that the majority of instruments in the DCART core set indeed fulfilled these prerequisites, which makes them useful as endpoints in clinical trials.
Chapter 3 focuses on measuring radiographic damage in AS. Radiographic damage was considered an important domain already by ASAS, but up to recently, an appropriately validated instrument to measure radiographic damage and progression of damage was lacking. In this chapter, three scoring methods to measure radiographic damage of the spine were compared with respect to the validation criteria proposed by OMERACT: 1) Truth, (*or.* does the instrument really measure what it is supposed to measure?); 2) Discrimination, (*or.* does the measure discriminate between groups of patients with different levels of severity?); and 3) Feasibility, (*or.* is it feasible to use this measure in the context of clinical trials?).

One scoring method, the modified Stoke Ankylosing Spondylitis Spinal Score (mSASSS), appeared to outreach the other two tested methods in terms of reliability (between readers) and sensitivity-to-change. Based on this study, the mSASSS was proposed as the method of choice to assess radiographic progression in AS clinical trials, and as such endorsed by OMERACT in 2004.

Chapter 4 focuses on a methodological issue pertaining to scoring radiographic progression in clinical trials: should radiographs be read "chronologically" (with known time order) or "paired" (with concealed time order)? The theoretical advantage of chronological reading (more signal picked up) can be a source of bias in clinical trials, because it may give the reader a direction to expected deterioration and the scores may be dependent on the magnitude of progression, which can be different in both treatment groups. At the other hand, paired reading may be at the cost of signal intensity. Signal intensity has always been a source of concern, since available data showed that measurable radiographic progression in patients with AS is slow. This study, however, showed that even with paired reading there was sufficient progression in a 2-year period to be picked up by the mSASSS. This finding is reassuring, since the methodology of measuring radiographic progression requested by registration authorities includes scoring with paired time order.

Chapter 5 addresses a clinical problem: "Is it really necessary to take radiographs of the spine to be informed about the extent of damage, or is it sufficient to clinically determine spinal mobility by one or more of the available spinal mobility instruments?" This study shows that spinal mobility and radiographic damage are expectedly correlated on the group level, but that concordance between spinal mobility measurements and radiographic damage are insufficient to assure an appropriate classification at the individual patient level. The conclusion of this study therefore was that spinal mobility measures could not replace radiography of the spine in order to get an impression about structural damage in patients with AS, which gives justice to the inclusion of the domain radiographic damage in the ASAS core set for DCART in addition to the domain spinal mobility.

Chapter 6 of this thesis treads upon a methodological by-path: Selection of patients for a clinical trial at baseline based on prediction of the outcome of interest. Usually, patients selected for a clinical trial have a relatively bad prognosis (*or.* a high propensity on an unfavorable outcome) because it is widely accepted that such a selection ameliorates the statistical demonstration of a treatment difference. This study shows that the extent to which treatment contrast is determined by baseline propensity is dependent on how a treatment reduces the risk of an unfavorable outcome: by absolute risk reduction or by relative risk reduction. Though relative risk reduction models (*eg.* 50% reduction of the risk irrespective of the baseline risk) are most prevalent in pharmacology, a few examples from rheumatoid arthritis clinical trials suggest an

absolute risk reduction model for two treatment strategies, which may have implications for the selection of patients for such trials.

Chapter 7 describes the results of a randomized controlled trial in which all methodological principles of scoring radiographic progression as mentioned before were applied. The trial compared two strategies of treatment with non-steroidal anti-inflammatory drugs (NSAIDs) in patients with AS: one strategy with the continuous use of NSAIDs, one strategy with on- demand-only use of NSAIDs. The primary outcome measure was radiographic progression after a 2-year follow up period. Radiographic damage was scored by mSASSS, as outlined before, with radiographs presented with concealed time order (paired reading). The study not only showed that progression was measurable in a considerable proportion of patients (confirming sensitivity to change of the method), but also that continuous NSAID use was associated with significantly less radiographic progression as compared to on-demand-only NSAID use (discriminatory capacity). These results showed that 2-year radiographic progression is a feasible endpoint in AS clinical trials, confirmed that the mSASSS is an appropriate instrument to assess radiographic progression, and shed new light on the potential of NSAIDs as drugs that may influence structural damage in AS.

How will the studies described in this thesis contribute to clinical science in the field of AS in the near future?

It is beyond any doubt that clinical outcome research in AS in the year 2005 stands at a different level as compared to a decade ago. As stated above, evaluation of TNF-blocking drugs and outcome research in AS have mutually catalyzed each other's development, reaching a state that approximates the level of development of rheumatoid arthritis. Putatively as a consequence of the speed of development, a few areas of outcome assessment in AS are relatively untouched. First, unlike the situation in RA, and despite many efforts in the past, it is still not known how to predict a favorable or an unfavorable long-term outcome of the disease at presentation. Second, whilst the efficacy of TNF-blocking drugs on signs and symptoms of the disease is undisputed, it is still unknown whether these drugs truly influence long-term outcome, including structural damage. Third, and related to some extent, many experts in the field consider the success of TNF-blocking drugs in AS, which is based on subjective improvements, insufficiently substantiated by improvements on objective instruments that are beyond the influence of patient's perceptions.
Prediction of long-term outcome in AS, drug effects on structural damage, and objective improvements by drugs will be major fields of research in the forthcoming years, which all require a relevant (set of) outcome measure(s). There are good methodological arguments to postulate that imaging - plain radiography of the spine to assess structural damage and magnetic resonance imaging (MRI) of the spine to assess inflammation

and to some extent structural damage - will play an increasingly important role in the assessment of various types of outcome in AS.

Radiography of the spine reveals structural damage, which is associated with spinal mobility and with function, and therefore is considered relevant to patients. What is still unknown is whether radiographic progression is determined by disease activity and/or inflammation, as in RA, but the analysis of databases of clinical trials and long-term observational studies will shed light on this relationship. A very important field of research will be the substantiation of inhibition of radiographic progression by TNF-blocking drugs. Most probably, the evaluation of a "structural damage claim" for such drugs will be unprecedented because normal trial designs do not suffice. It is widely considered unethical to perform placebo-controlled clinical trials in which patients are refrained from effective therapy for a period of at least 2 years, required to obtain a sufficiently large treatment contrast in radiographic progression. Technical developments in the processing of radiographs (*eg.* digitization) make it possible to match radiographs of actively treated patients with radiographs of patients from untreated cohorts in an indistinguishable manner, and offer these radiographs to readers for scoring in a concealed time order. Obviously, such an experimental set-up lacks the advantage of prognostic similarity of treatment groups, which is inherent to randomization. But differences between artificial treatment groups are only relevant in so far they determine radiographic progression. Unlike the situation in RA, up to now, thorough analysis with multiple demographic, clinical, laboratory and genetic variables has not revealed a single variable that contributes to explaining 2-year radiographic progression. It is therefore to be expected that – given the impossibility of a 2-year randomized controlled trial - a controlled cohort design, in which radiographic progression of actively treated AS patients is compared with radiographic progression of not concurrently treated control patients, will reveal useful information with regard to the potential of TNF-blocking drugs to inhibit progression of structural damage, and to fill in lacking information pertaining to objective outcome domains.

MRI of the spine is an imaging modality with promising prospects. MRI techniques visualize inflammation of the vertebrae and adjacent structures, as well as structural damage such as the formation of syndesmophytes and vertebral erosions. Preliminary data have shown that "MRI activity" of the spine can be suppressed by TNF-blocking drugs with a time interval of 6 weeks, whilst being unchanged in placebo treated patients. The actual place of MRI as a measurement instrument in AS clinical trials, and later on in clinical practice, is dependent on how accurately MRI inflammation reflects signs and symptoms of AS, function and quality of life, as expressed by patients, and how accurately MRI inflammation predicts impairment of spinal mobility and structural damage later on. An international working group under the umbrella of ASAS and OMERACT, and consisting of experts in the field of imaging, is now pursuing a thorough validation program in order to establish the value of MRI of the spine as an outcome instrument for AS. It is to be expected that "MRI-activity" will become important in the decision whether a patient with AS has active disease, whether this patient has an

unfavorable prognosis, and whether treatment with TNF-blocking drugs should be started. Other fields in development are biomarker research, genomics and proteomics which are also applied to AS. MRI in combination with plain radiography, together with the achievements of these new research fields, may help elucidating pathophysiological mechanisms underlying AS.

In summary, AS is a popular research field nowadays. A better understanding of the disease mechanisms, of its course and impact, and of the various effects of drugs is all within the scope of currently planned research. A high standard of outcome assessment is crucial in this process, and the studies described in this thesis may have added to this.

# Samenvatting in perspectief

# Samenvatting in perspectief

De in dit proefschrift beschreven studies bevatten belangrijke aspecten van *outcome assessment* (het beoordelen van de lange termijn uitkomst) van de ziekte van Bechterew. Het onderzoek op het gebied van de ziekte van Bechterew heeft zich tientallen jaren voornamelijk gericht op de pathofysiologische mechanismen van deze aandoening, en niet zozeer op *outcome assessment*. Dit laatste werd als onnodig beschouwd omdat er immers geen medicijnen beschikbaar waren die het ziektebeloop positief konden beïnvloeden. Daarnaast werd ook bijna geen onderzoek gedaan met medicamenten.
Twee opeenvolgende ontwikkelingen hebben dit beeld drastisch veranderd. Allereerst vond in 1995 de oprichting van de internationale werkgroep "ASsessment in Ankylosing Spondylitis (ASAS)" plaats te Amsterdam. Deze internationale groep van deskundigen definieerde de aspecten van de ziekte van Bechterew die gemeten dienden te worden voor verschillende situaties, bijvoorbeeld een klinische trial, of de dagelijkse praktijk. Ook selecteerde zij de bijbehorende meetinstrumenten. Deze ASAS activiteiten werden in 1998 bekrachtigd door OMERACT, het wereldwijde initiatief op het gebied van het ontwikkelen van uitkomstmaten bij reumatische ziekten.
Een tweede belangrijke ontwikkeling was de ontdekking van biologische medicijnen die in staat waren om tumor necrosis factor alpha te remmen, de zogeheten anti-TNF middelen. Tumor necrosis factor alpha is een stof die deel uitmaakt van een ontstekingscascade bij chronische ontstekingsziekten. De introductie van deze middelen bij de behandeling van patiënten met reumatoïde artritis liet spectaculaire resultaten zien. Op kleinere schaal werd er ook geëxperimenteerd met patienten met de ziekte van Crohn (een ziekte die geassocieerd is met de ziekte van Bechterew) en bij patiënten met de ziekte van Bechterew. Deze eerste resultaten waren positief. De farmaceutische industrie wilde dan ook de anti-TNF middelen laten registreren voor de ziekte van Bechterew, waarvoor grote klinische trials noodzakelijk zijn, met geschikte meetinstrumenten. Het kan rustig worden gesteld dat de toepassing van anti-TNF middelen bij de ziekte van Bechterew heeft gezorgd voor een enorme stimulans voor het onderzoek op het gebied van *outcome assessment* bij de ziekte van Bechterew. Deze ontwikkelingen zijn een rechtvaardiging voor het wetenschappelijke werk zoals samengevat in dit proefschrift, en hebben tevens de onderzoeksagenda voor het Bechterew onderzoek voor de komende jaren bepaald.

Met de komst van de anti-TNF middelen lijkt het erop dat er medicatie beschikbaar is gekomen waarmee het ziektebeloop fundamenteel kan worden beïnvloed. Of dit ook daadwerkelijk het geval is zal door onderzoek in klinische trials moeten worden uitgezocht. Voor dit doel heeft de ASAS een speciale set meetinstrumenten ontwikkeld, de DCART core set. In hoofdstuk 2 wordt een studie beschreven waarin de gevoeligheid en het onderscheidend vermogen van de instrumenten behorend tot deze set onderzocht wordt. Hierbij wordt gebruik gemaakt van de resultaten van een van de

eerste trials waarin anti-TNF medicatie bij Bechterew patiënten werd getest. Het bleek dat het merendeel van de geselecteerde instrumenten over een voldoende gevoeligheid en onderscheidend vermogen beschikten, hetgeen ze tot bruikbare instrumenten maakt voor het beoordelen van klinische trials.

Hoofdstuk 3 richt zich op het meten van radiologische schade van de ziekte van Bechterew. Door de ASAS werkgroep werd radiologische schade als een belangrijke uitkomst gezien, echter een gevalideerd en geschikt instrument om radiologische schade en het voortschrijden van deze schade (progressie) te meten, ontbrak. In dit hoofdstuk worden drie scoringsmethodes, die de radiologische schade van de wervelkolom meten, vergeleken op basis van validatie-criteria zoals voorgesteld door de OMERACT. Deze criteria zijn als volgt: 1) Meet het instrument daadwerkelijk wat het veronderstelt te meten? 2) Kan het instrument onderscheid maken tussen groepen patiënten met een verschillende ernst van de aandoening? . 3) Is het praktisch haalbaar om het instrument toe te passen in de context van een klinische trial? Deze drie criteria worden in het OMERTACT filter aangeduid met de volgende Engelse termen; truth, discrimination and feasibility.
Eén scoringsmethode, de gemodificeerde Stoke Ankylosing Spondilitis Spinal Score (mSASSS), bleek beter dan de andere twee met betrekking tot betrouwbaarheid (dat wil zeggen: de hoogste mate van overeenkomst tussen verschillende gebruikers van de methode) en de mSASSS bleek het meest gevoelig voor veranderingen. Gebaseerd op deze studie werd de mSASSS voorgesteld als de methode van eerste keus om radiologische progressie in klinische trials bij Bechterew patiënten te beoordelen. Dit voorstel werd in 2004 door OMERACT bekrachtigd.

Hoofdstuk 4 richt zich op de methodologie van de scorings volgorde van röntgenfoto's om radiologische progressie te beoordelen. De vraag hierbij is of de foto's van een patiënt gescoord moeten worden in een chronologische, dus bekende volgorde, dan wel in een willekeurige volgorde. Het theoretische voordeel van het scoren in chronologische volgorde is dat er een grotere verandering wordt gezien. Maar dit voordeel is ook gelijk een nadeel. Het bekend zijn van de volgorde kan er immers toe leiden dat men geneigd is meer schade te zien dan er daadwerkelijk is, omdat men meer schade verwacht. Het scoren van de foto's in willekeurige volgorde heeft dit nadeel niet. Maar men is geneigd behoudender te scoren, en het gevaar bestaat dat een daadwerkelijk bestaand verschil tussen twee verschillende groepen niet wordt gezien. Op die manier gaat het scoren met een willekeurige volgorde ten koste van signaalintensiteit. De signaalintensiteit is altijd al een bron van zorg geweest omdat de progressie bij de ziekte van Bechterew zeer langzaam verloopt. De studie beschreven in dit hoofdstuk laat echter zien dat zelfs met het scoren van de foto's in willekeurige volgorde middels de mSASSS, er meetbare progressie optreedt gedurende een follow up duur van 2 jaar.

In hoofdstuk 5 komt de volgende klinische vraag aan de orde: " Is het werkelijk nodig om röntgenfoto's van de wervelkolom te nemen om geïnformeerd te worden over de mate van schade, of is het voldoende om klinisch de beweeglijkheid van de wervelkolom te bepalen met behulp van één of meer meetinstrumenten? Deze studie laat zien dat er zoals verwacht een correlatie bestaat tussen spinale mobiliteit en radiologische schade, maar dat de overeenkomst tussen deze spinale mobiliteit en radiologische schade onvoldoende is om de individuele patiënt correct te classificeren. De conclusie van deze studie is dan ook dat het meten van de beweeglijkheid van de wervelkolom niet als vervanging kan dienen voor röntgenfoto's om een indruk te krijgen van de structurele schade bij Bechterew patiënten. Deze conclusie rechtvaardigt de inclusie van het beoordelen van radiologische schade naast het beoordelen van de spinale mobiliteit in de DC-ART core set van de ASAS.

Hoofdstuk 6 van dit proefschrift beschrijft een uitstapje op een methodologisch zijpad: selectie van patiënten voor een klinische trial. Normaal gesproken worden patiënten geselecteerd die een slechte prognose hebben of te wel een hoge kans op een ongunstig resultaat. Dit omdat een dergelijke selectie de kans het grootst maakt dat statistisch kan worden aangetoond dat een behandeling daadwerkelijk effect heeft. De in dit hoofdstuk beschreven studie laat zien dat de mate waarin selectie het behandelingscontrast beïnvloedt afhangt van de manier waarop de behandeling het risico van de ongunstige uitkomst reduceert; door absolute dan wel relatieve risico reductie. Hoewel relatieve risico reductie modellen (bijvoorbeeld 50% minder kans op de ongunstige uitkomst aan het eind van de behandeling ten opzichte van het begin risico) het meest voorkomen binnen de farmacologie, suggereren een aantal voorbeelden afkomstig uit trials met reumatoïde artritis patiënten dat een absoluut risico reductie model werkzaam is. Dit zou implicaties kunnen hebben voor de selectie van patiënten voor klinische trials.

Hoofdstuk 7 beschrijft de resultaten van een gerandomiseerde gecontroleerde trial waarin alle eerder beschreven methodologische principes worden toegepast. In de trial worden twee behandelingsstrategieën met non-steroidale anti-inflammatoire medicijnen (NSAIDs) bij Bechterew patiënten vergeleken. Bij de eerste behandelstrategie gebruikt de patiënt voortdurend een NSAID en bij de tweede strategie gebruikt de patiënt alleen een NSAID als hij/zij klachten heeft, dan wel symptomen bemerkt ("naar behoefte"). Als primaire uitkomstmaat werd gekeken naar de radiologische progressie over 2 jaar. De radiologische progressie werd gescoord middels de mSASSS waarbij de foto's in willekeurige volgorde werden bekeken. Deze studie liet niet alleen zien dat er bij een aanzienlijk deel van de patiënten inderdaad progressie zichtbaar was maar ook dat het continue gebruik van NSAIDs geassocieerd was met een significant lagere radiologische progressie als werd vergeleken met de patiënten die een NSAID gebruikten naar behoefte. Deze resultaten bevestigen dat radiologische progressie over 2 jaar een praktisch haalbare uitkomstmaat is in trials met Bechterew patiënten. Tevens

bevestigen deze resultaten dat de mSASSS een geschikt instrument is voor het meten van radiologische progressie en werpen zij nieuw licht op de rol die NSAIDs spelen in de behandeling van Bechterew patiënten.

Hoe kunnen de studies die zijn beschreven in dit proefschrift bijdragen aan het klinisch wetenschappelijk onderzoek naar de ziekte van Bechterew in de nabije toekomst?

Het is zonder enige twijfel een feit dat het klinisch onderzoek met betrekking tot *outcome assessment* van de ziekte van Bechterew van een geheel ander nivo is dan een decennium eerder. Zoals reeds eerder vermeld hebben de komst van anti-TNF middelen en het onderzoek van *outcome assessment* bij Bechterew elkaars ontwikkeling gekatalyseerd. Dit heeft voor *outcome assessment* geleid tot een status die ongeveer gelijk is aan die van *outcome assessment* in het onderzoeksveld van reumatoïde artritis, een veld dat tot voor kort een grote voorsprong had. Door de enorme snelheid waarmee de ontwikkelingen zich binnen het Bechterew onderzoek hebben voltrokken zijn er een aantal gebieden die relatief weinig aandacht hebben gehad. Ten eerste is het nog steeds niet bekend, ondanks diverse pogingen, hoe bij vroege ziekte het langere termijn beloop voorspeld kan worden, dit in tegenstelling tot de situatie bij reumatoïde artritis. Ten tweede: het effect van anti-TNF middelen op klachten en symptomen van Bechterew is onomstreden, maar het is nog steeds onbekend of deze middelen de lange termijn uitkomst beïnvloeden, inclusief structurele schade aan de wervelkolom. Als derde punt sluit hierbij aan dat een aantal experts het succes van anti-TNF blokkers bij de ziekte van Bechterew beschouwen als een succes dat gebaseerd is op subjectieve symptomen, en niet op objectieve instrumenten (instrumenten die niet kunnen worden beïnvloed door de perceptie van de patiënt).
Deze drie punten, het voorspellen van de lange termijn uitkomst van Bechterew, de effecten van medicatie op structurele schade, en objectieve verbeteringen door medicatie, zullen de komende jaren belangrijke onderwerpen zijn in het Bechterew onderzoek. Al deze zaken vereisen relevante uitkomstmaten. Er zijn goede methodologische argumenten om te stellen dat de conventionele röntgenfoto van de wervelkolom om structurele schade te beoordelen en een MRI van de wervelkolom om de ontsteking te beoordelen (en tot op bepaalde hoogte ook de structurele schade), een steeds belangrijker rol zullen gaan spelen bij het beoordelen van de uitkomst van de ziekte van Bechterew.
Radiologie van de wervelkolom laat structurele schade zien, die geassocieerd is met spinale beweeglijkheid en met functioneren. Door deze associatie wordt radiologie beschouwd als relevant voor de patiënt. Het is nog steeds onbekend of radiologische progressie bepaald wordt door ziekte activiteit en/of ontsteking, zoals bij reumatoïde artritis het geval is. Analyses van gegevens van klinische trials en langdurig observationeel onderzoek zullen hopelijk meer licht werpen op deze relatie.
Een zeer belangrijk onderwerp van onderzoek zal zijn het bevestigen van het remmende effect van de anti-TNF middelen op radiologische progressie bij Bechterew

patiënten. Hierbij doet zich echter een ethisch probleem voor. Het is onverantwoord om een placebo gecontroleerde trial te doen gedurende 2 jaar, de tijd die nodig is om een verschil in radiologische progressie aan te tonen, en daarmee de patiënten behorend tot de placebo groep een therapie die effectief is om de klinische verschijnselen te bestrijden, te onthouden.

Technische ontwikkelingen, zoals het digitaliseren van oude röntgenfoto's, maken het mogelijk om röntgenfoto's van patiënten die met anti-TNF middelen zijn behandeld te matchen met reeds bestaande röntgenfoto's van patiënten die niet behandeld zijn en deze aan te bieden om te scoren en te vergelijken. Omdat in een dergelijke experimentele opzet geen sprake is van randomisatie (toevalsverdeling) zijn de behandelgroepen niet per definitie gelijk qua prognose van de ziekte. Echter een verschil in prognostische gelijkheid is alleen relevant als het gaat om factoren die de uitkomst, in dit geval radiologische progressie, beïnvloeden. En voor Bechterew, in tegenstelling tot reumatoïde artritis, is ondanks grondige analyse van meerdere demografische, klinische, laboratorium en genetische variabelen, er geen enkele variabele aan het licht gekomen die als prognostische factor aangemerkt kan worden. Vandaar dat kan worden verwacht dat een experiment als hierboven beschreven bruikbare informatie zal opleveren met betrekking tot het effect van anti-TNF middelen op radiologische schade.

MRI van de wervelkolom is een beeldvormende techniek met veelbelovende verwachtingen. Middels MRI kan ontsteking van de wervels en van de aanliggende structuren zichtbaar gemaakt worden, evenals structurele schade, zoals syndesmofyten en vertebrale erosies. Voorlopige gegevens hebben laten zien dat ontstekingsactiviteit zoals door MRI wordt weergegeven, kan worden onderdrukt door het gedurende 6 weken gebruiken van anti-TNF middelen, terwijl deze ontstekingsactiviteit in met placebo behandelde patiënten onveranderd was gebleven. De plaats van MRI als een meetinstrument voor Bechterew in klinische trials, en later in de klinische praktijk, is afhankelijk van de vraag hoe accuraat de ontstekingsactiviteit zoals gezien middels MRI, de klachten en symptomen, functie en kwaliteit van leven van Bechterew patiënten weergeeft, en hoe accuraat de ontstekingsactiviteit zoals waargenomen op MRI de spinale mobiliteit en structurele schade voorspelt. Een internationale werkgroep bestaand uit experts op het gebied van beeldvorming is nu onder de paraplu van ASAS en OMERACT bezig om de waarde van MRI van de wervelkolom als meetinstrument vast te stellen. Naar verwachting zal ontstekingsactiviteit zoals door MRI waargenomen een belangrijk item worden in de besluitvorming of een Bechterew patiënt een actief ziektebeloop heeft, of deze patiënt een ongunstige prognose heeft en of behandeling met anti-TNF gestart moet worden. Op het gebied van onderzoek naar biomarkers, genen en eiwitten zijn er ontwikkelingen gaande die ook hun toepassing in het Bechterew onderzoek zullen vinden. MRI in combinatie met conventionele radiologie en de resultaten van de hier bovengenoemde onderzoeksvelden kunnen behulpzaam zijn bij het ontrafelen van het pathofysiologisch mechanisme dat ten grondslag ligt aan de ziekte van Bechterew.

Samengevat kan worden gesteld dat het onderzoek op het gebied van Bechterew momenteel volop in de belangstelling staat. Een beter begrip van de ziekte-mechanismen, van het ziektebeloop en de gevolgen, en van de vele effecten van behandeling met medicijnen zijn allen binnen het bereik van het huidige lopende en geplande onderzoek. In dit proces is een hoge standaard met betrekking tot *outcome assessment* cruciaal, en de studies beschreven in dit proefschrift hebben daar zo mogelijk aan bijgedragen.

# Dankwoord

# Dankwoord

Een proefschrift schrijven kent vele lastige facetten waarvan het schrijven van een dankwoord er één is, immers dit is het meest en ik denk ook het meest kritisch gelezen hoofdstuk van het boekje. Anderzijds is het ook het leukste gedeelte om te schrijven, immers al nadenkend over de inhoud passeren vele personen mijn gedachten en het is prettig om mijn erkentelijkheid jegens hen te mogen verwoorden voor een relatief breed lezerspubliek.

Allereerst mijn promotoren en co-promotor; Sjef van der Linden, Désirée van der Heijde en Robert Landewé. Beste Sjef, van een afstand, maar toch betrokken, heb je jouw bijdrage geleverd aan de realisatie van dit proefschrift. Van je kritische commentaar op de verschillende manuscripten heb ik veel geleerd, dank hiervoor. De 'dagelijkse' begeleiding van dit proefschrift is in handen geweest van Désirée en Robert. Onze bijeenkomsten om te discussiëren over de inhoud van dit proefschrift heb ik altijd als zeer inspirerend en motiverend ervaren. Door de bezoeken aan buitenlandse congressen leerde ik de onderzoekswereld rondom Bechterew kennen en ook de vooraanstaande rol die de Maastrichtse onderzoeksgroep daarin speelt. Ik waardeer het zeer dat jullie mij de kans hebben gegeven om daaraan ook mijn bijdrage te mogen leveren. Désirée en Robert, ik bewonder jullie om jullie tomeloze inzet en bevlogenheid voor wetenschappelijk onderzoek en ik ben jullie dankbaar voor het enthousiasmerende effect daarvan. Daarnaast heb ik jullie betrokkenheid en aandacht voor andere hoogte- en dieptepunten in mijn leven zeer gewaardeerd.

Voor mij is het belangrijk om naast inhoudelijk interessant werk ook een prettige werkomgeving te hebben. Hieraan is meer dan voldaan, grotendeels dankzij een aantal kamergenoten die ik graag wil noemen.
Jolanda Brauer, tijdens het eerste jaar van mijn aanstelling had ik genoegen om met jou een kamer te mogen delen, dit was meerdere malen een waar feest.
Karin Bruynesteyn, met jou als kamergenoot werd het leven een stuk serieuzer. We belanden regelmatig in ingewikkelde discussies over abstracte zaken waarin ik mezelf soms verbaasde dat ik je kon volgen. Ik heb jouw aanwezigheid als zeer stimulerend ervaren. Ik bewonder je passie voor onderzoek en je perfectionistische manier van werken. Collega's zijn we inmiddels niet meer, vriendinnen nog wel en ik hoop dat dit zo mag blijven. Het geeft me een goed gevoel jou als paranimf aan mijn zijde te hebben.
Liesbeth Heuft, altijd samen op zoek naar manieren om onderzoeksfrustraties uit het lijf te jagen en de innerlijke balans te herstellen. Na onmogelijke steps-stapjes en complexe yoga poses bleek uiteindelijk Sleepy Time thee nog het meest effectief.
Simone Gorter, in de laatste fase van mijn onderzoek was jij mijn kamergenoot en ik heb genoten van je gezelligheid die me al voor mijn vertrek uit Maastricht met weemoed vervulde.

Naast kamergenoten waren er meer collega onderzoekers die ik zeer erkentelijk ben voor de samenwerking en hun gezelligheid, dank jullie wel; Astrid van Tubergen, Guy Schulpen, Erik de Klerk, Debbie Vosse en Annelies Boonen.

Een speciaal woord van dank gaat uit naar Anneke Spoorenberg. Anneke, al snel na mijn aanstelling ben je vertrokken naar het hoge Noorden. Daar mijn onderzoek voor een groot deel voort borduurt op jouw onderzoekswerkzaamheden hebben we contact gehouden en ben je me altijd zeer behulpzaam geweest. Zo heb je me o.a. enorm geholpen met het leren scoren van röntgenfoto's. Jouw betrokkenheid en interesse heb ik zeer gewaardeerd.

Naast de onderzoekers ben ik ook de overige leden van de werkgroep reumatologie erkentelijk voor het prettige werkklimaat. Reumatologen, arts-assistenten, en onderzoeksassistenten. Met name wil ik ook de secretaresses Marianne Curfs, Yolanda Soons en Femke Hoekstra bedanken voor de ondersteuning en persoonlijke interesse.

De hoofdstukken in dit proefschrift zijn niet alleen van mijn hand. Dankzij de bijdragen van een groot aantal medeauteurs hebben de bijbehorende artikelen inmiddels hun weg gevonden naar de diverse reumatologische tijdschriften. Special thanks to all co-authors who contributed to the several papers. It has been a privilege to work with all of you.

Waardering voor de persoon Tiny Wouters ontbreekt in geen enkel proefschrift voortkomend uit de vakgroep Interne Geneeskunde. Zoals ik inmiddels heb mogen ondervinden is dit geheel terecht. Tiny, jouw enthousiasme en enorme inzet om een keurig verzorgd eindresultaat te produceren is onovertroffen.

Ook dank aan alle vrienden, familie en bekenden die hun interesse en belangstelling toonden. Een aantal personen wil ik met name noemen. Allereerst Nettie van der Meer. Na onze gezamenlijke start in Maastricht vind ik het fijn dat je ook nu weer, op één van de hoogtepunten van mijn leven, mij ter zijde wilt staan, dit keer als paranimf.
Ten tweede Arie, Bea, Vincent en Noémi van Duijvenbode. Officieel familie van de 'kouwe kant', maar ik heb altijd een warme belangstelling en interesse voor mijn 'scriptie' ervaren, dank!

Lieve pap en mam, dank voor het warme nest wat jullie hebben geboden en bieden. Dit is de basis onder wie ik ben en wat ik doe. Jullie onvoorwaardelijke liefde en vertrouwen hebben veel mogelijk gemaakt.
Lieve Maurice, mijn kleine grote broertje, naast mijn dankbaarheid voor wie jij bent, ben ik je vooral dankbaar dat je er bent.
Lieve Jeroen, je kunt niets zeker weten en alles gaat voorbij, maar met een aan zekerheid grenzende waarschijnlijkheid durf ik toch wel te stellen dat al mijn later is met jou, en die wetenschap maakt alles mogelijk.

# Curriculum vitae

# Curriculum vitae

Astrid Wanders werd geboren op 13 februari 1973 te Eerbeek (Gld). In 1991 behaalde zij haar VWO diploma aan de Heemgaard te Apeldoorn. Vervolgens zou ze graag gestart zijn met de studie Geneeskunde maar werd helaas uitgeloot en begon aan de studie Gezondheidswetenschappen aan de universiteit van Maastricht. In 1992 behaalde zij het propaedeuse diploma en vervolgde met de bovenbouw studie Milieugezondheidkunde, differentiatie Onderzoek. Hiervan werd het doctoraal diploma in 1995 behaald. In 1994 was het lot haar gunstig gezind en kon zij starten met de studie Geneeskunde waarvan in 2000 het artsexamen werd behaald. Op 1 juni 2000 trad zij als arts-onderzoeker voor een periode van 3,5 jaar in dienst van de werkgroep reumatologie van het academisch ziekenhuis Maastricht. In het eerste jaar van haar aanstelling hield zij zich bezig met het uitvoeren van een fase II geneesmiddelenonderzoek in een drietal ziekenhuizen in Zuid-Limburg. De overige 2,5 jaar van haar aanstelling werd besteed aan het onderzoek zoals gepresenteerd in dit proefschrift. In december 2003 begon zij aan de vooropleiding Interne Geneeskunde in het Meander Medisch Centrum te Amersfoort in het kader van de opleiding tot reumatoloog. Al snel werd het haar duidelijk dat het ingezette traject niet het juiste was. Na een periode van her-oriëntatie begon zij in maart 2004 als AGNIO kinder- en jeugdpsychiatrie op de polikliniek de Riethorst van GGZ Meerkanten te Ermelo. Sinds augustus 2004 combineert zij de polikliniche zorg voor kinderen met de deeltijdbehandeling voor adolescenten. Het plezier en de voldoening die deze werkzaamheden haar boden leidde tot een definitieve keuze voor de psychiatrie. Vanaf september 2005 zal zij starten met de opleiding tot psychiater die verbonden is aan GGZ Meerkanten te Ermelo (opleider dr. H. van Megen). Naast dat zij in 2004 haar ware liefde binnen de geneeskunde ontmoette, werd door middel van het huwelijk de relatie met haar andere ware liefde bezegeld. Sinds 4 september 2004 is zij zeer gelukkig getrouwd met Jeroen van Duijvenbode.